A Study of Notions of Participation and Discourse in Argument Structure Realisation

Brian Murphy

PhD Computer Science

2007

Supervisor: Dr. Carl Vogel

Declaration

I hereby declare that this thesis is entirely my own work and that it has not been submitted as an exercise for a degree at any other university. I give permission to the Library to lend or copy this thesis upon request.

Brian Murphy, November 2007

Acknowledgements

First and foremost I would like to thank my supervisor Carl. He was unfailingly generous with his time, and always provided hands-on support as this work developed. Working with him was a truly enjoyable and mind-broadening experience, which led me along winding path from computational linguistics to cognitive science. I would also like to thank the examiners Tim Fernando and particularly Ellen Bard for their detailed and insightful comments – this work is much improved as a result.

General academic thanks to all my colleagues at the Computational Linguistics Group, for chats and for always being willing pilot participants. They and my undergraduate students also have to be thanked for keeping me sane over the years of study.

Thanks to my mother and father for getting me here in the first place. And thank you Clare for your endless patience, love and support – not to mention the copious quantities of great food that literally fuelled this thesis. A final special thanks to my son, whose impending arrival provided an extra special motivation during revisions.

This work could not have been carried out without the support of a generous teaching fellowship (TCD Senior Lecturer's Broad Curriculum Fellowship) from the Graduate Studies Office and the Department of Computer Science, support for fees also from the Graduate Studies Office, and an initial grant from Enterprise Ireland.

Contents

List of Figures

List of Tables

Abstract

There is a lack of consensus on the nature of the process of linking; that is, how the conceptual form of thoughts determines the form of the spoken expressions used to describe them. Prominent theories of argument structure realisation include ones based on fixed inventories of participant types (semantic roles, Fillmore, 1968), detailed analyses of the lexical semantics of verbs (predicate decompositions, Jackendoff, 1972; Rappaport and Levin, 1988), characterisations of argument structure positions in terms of prototype semantics (proto-roles, Dowty, 1991), and factors of pragmatics and perspective (Kuno, 1976; Croft, 1991). Given the wide range of views, this study attempts to investigate empirically which semantic and pragmatic factors are the primary drivers of argument structure realisation, without making detailed assumptions about processing models or linguistic representations. The first set of experiments carried out gathered large numbers of acceptability judgements from many non-linguist native-speaker participants over the internet, to quantify the relative acceptability of minimal pairs of diathesis variants. Film script extracts were used to provide authentic materials and contextualisation, and established methodologies from psycholinguistics were employed (Bard et al., 1996; Cowart, 1997). The diatheses examined were the English dative and benefactive constructions; the English and German passive constructions; and the Chinese *ba* and *bei* constructions. Effects of aspects of involvement and of discourse structuring were found in all three languages, and crucially the magnitude of these effects was comparable to those elicited by clear syntactic violations. A follow-on corpus study on the British National Corpus gave qualified support to some of these effects. An annotation exercise among linguists found a reasonable degree of consensus on the use of traditional role inventories, but indicated that a coarser grained set of roles might be more appropriate. Overall, the results of these experiments suggested that strict modularity of processing (see e.g. Chomsky, 1986; Bock and Levelt, 1994) may not be a safe assumption; that competition between candidate realisations operates at the clause level; and that the determining factors in argument structure realisation are specific properties of the situation being described in an utterance, rather than default properties of the event or verb in question. On this basis a simple numerical model of argument structure production in English was constructed, integrating elements of Dowty's proto-roles with the pragmatic notions of information structure and topical saliency. This proof-of-concept model explains much of the linking patterns described in the literature, and provides qualified predictions of judged acceptability. This leads to speculation that such models have the potential to account for argument linking in a range of languages, nominative and otherwise.

Chapter 1

Introduction

> "And therefore, for any man to go about to add to their number, will be but like the inventing of a Disease, for which he can expect but little thanks from the World."
>
> Wilkins (1668, p.13) speaking on the prospect of new artificial languages

The work presented in this thesis is a contribution to the large body of work on the nature of linking – that is, how the conceptual form of situations determines the form of the expressions used to describe them. Rather than proposing an intricate representational model directly on the basis of an articulated theory (an artificial language of sorts), the approach adopted here is to row back on many of the assumptions made about the relationship between syntax, semantics and discourse. A series of empirical studies attempt to identify some of the properties of events and their participants that most strongly determine argument structure realisation in English. Some of these studies are then replicated in German and Chinese to test their language generality. No assumptions of modularity are made[1], and features of semantics and discourse are equally entertained. In particular, this work attempts to explain as much of linking as possible using general principles of grammar, rather than searching for explanations in the lexical semantics of individual verbs.

This chapter starts with an overview of different approaches that linguistics has taken to the question of argument linking, addressing the tension between language-general and verb-specific patterns: general phrase structure rules and transformations; lexicon-based listing approaches; semantic role category schemas; decomposition of verb lexical semantics; prototype characterisations of argument structure positions; and approaches that look to elements of discourse structure and pragmatics to explain argument structure realisation. Three subsections then follow to examine the minimal assumptions made in this work: on the nature of semantics and syntax; equivalence of meaning in paraphrases and diatheses (i.e. alternations); and the relationship between theoretical grammaticality and judged acceptability. The chapter then concludes with an outline of the structure of this document.

The mapping between meaning and argument structure is certainly not a simple one, even within a single language. Across languages it varies widely, the opposition of nominative and ergative languages being the most dramatic one. Early generative accounts of argument structure realisation (Chomsky, 1957) provided a simple account of how argument structure patterns for a language were specified by universal phrase structure rules – a monotransitive sentence such as 1.1 could

[1] Neither in the global sense of Fodor (1983), nor in the language internal sense of Chomsky (e.g. 1986) or Bock and Levelt (1994).

be generated by the rules in 1.2. Diatheses such as the passive (1.3) could be accounted for by transformational rules (1.4).

(1.1) [$_S$ [$_{NP}$ John] [$_{VP}$loves [$_{NP}$Mary]]]

(1.2) a. S → NP VP

b. VP → V NP

(1.3) Mary is loved by John

(1.4) NP_1 V NP_2 → NP_2 be V by NP_1

However it is clear that additional constraints apply. For example a large class of predicative and stative verbs that take a single noun-phrase argument nevertheless resist passivisation (consider *be, resemble, seem, weigh*), and there are many semantic constraints on admissibility of noun-phrases as verb arguments (consider **Golf played John* and **Send Rome that package*).[2]

Subsequent work in the generative semantics tradition (Katz and Fodor, 1963; Katz and Postal, 1964) suggested that verbs specified semantic features of admissible arguments by so-called selectional restrictions (e.g. the subject of *play* should be animate; the object of *phone* should be animate), but again it is easy to find counterexamples:

(1.5) I phoned [the hospital]$_{inanimate}$ again after Christmas.[3]

Chomsky later suggested that the argument structure of basic clauses is projected from the lexical entry of the verb in question (Chomsky, 1981). Large-scale lexica have been constructed to specify the linking patterns between semantics and syntax for individual verbs (e.g. Fillmore, 1982; Palmer, 2000; Dorr, 2001) but these can miss important generalisation of the syntax semantics mapping (Levin and Rappaport Hovav, 2005, pp.20-24). Argument structures are not entirely arbitrary. For example, a wide class of verbs that denote tangible or metaphorical transfer (e.g. *give, send*; *tell, teach*) can undergo the dative diathesis (Levin, 1993). Some pairs of near-synonymous verbs, such as *raise/rise*, *fell/fall*, *set/sit* and *lay/lie* contrast an explicitly causative event (Agent-Patient) with one in which causation is either implicit or absent. A sentence of the first type, e.g. *X raised Y*, is often synonymous with a causative construction containing the counterpart from the second set, e.g. *X made Y rise* (though in this case the causation may not be as immediate). English also has many unaccusative verbs such as *break* or *melt* that can appear with the transitive frame Agent–Verb–Patient also.[4] But this pattern is not universally productive, as verbs of similar meaning may variously require a transitive construction (e.g. *destroy*), or an unaccusative (e.g. *collapse*) form, or accept either (e.g. *break down*) (1.6). Even such requirements are subject to exceptions – in (1.7) *collapse* is used transitively.[5]

(1.6) a. The hordes broke down/destroyed/*collapsed the city walls

b. The city walls broke down/*destroyed/collapsed

[2]The hash symbol '*' is used to mark sentences that are syntactically illformed. See the appendix (page 183) for a list of marks and other notations used. The use of such informal notation of acceptability is discussed in section 1.3.

[3]Radio Programme Summary. [http://www.rte.ie/radio1/liveline/1020814.html – viewed August 2007]

[4]Here unaccusative refers to any monotransitive verb with a patient-like subject, without making any commitment to the Chomskyan conception of such verbs having an underlying object.

[5]The *fold-up* sense of *collapse* is routinely transitive; e.g. *Clare collapsed the fold-up picnic table*.

(1.7) ... a truck filled with high explosives ... drove into the lobby ... and detonated, **collapsing all four floors** of that building down upon itself.[6]

Neither can listing approaches account for novel argument structures, such as the use of *gold-watch* as a verb below (1.8), or the transitive use of the verb *sneeze* to express caused motion (1.9).

(1.8) G.W. Dickinson, ex-superintendent of the Montana Union, has been gold-watched by the employees of the company.[7]

(1.9) Sally sneezed the napkin off the table[8]

Construction Grammar accounts (Fillmore and Kay, 1987; Goldberg, 1995, 2006) have suggested that argument structures should be seen as lexical entities in themselves, which can be polysemous, in the same way that words are: central senses are related to the prototypical use of an argument structure, and other senses can be derived by relaxing the central definition (see Goldberg, 1995, chapter 3) or by metaphorical extension (e.g. compare *give someone a toy* and *give someone a headache*). For example the dative construction (1.10) is viewed by Goldberg as having six senses (1995, p.75), with the five non-central senses inheriting most of their structure from the central sense. But clearly, such an approach is similarly subject to the problems of delineating word senses in the lexicon, and in deciding which sense a particular utterance expresses (see for example Robins, 1987; Geeraerts, 1993; Kilgarriff, 1997a,b).[9]

(1.10) a. 'X CAUSES Y TO RECEIVE Z' (central sense)
Example: Joe gave Sally the ball.

b. Conditions of satisfaction imply 'X CAUSES Y TO RECEIVE Z'
Example: Joe promised Bob a car.

c. 'X CAUSES Y NOT TO RECEIVE Z'
Example: Joe refused Bob a cookie.

d. 'X ACTS TO CAUSE Y TO RECEIVE Z at some future point in time'
Example: Joe bequeathed Bob a fortune.

e. 'X ENABLES Y TO RECEIVE Z'
Example: Joe permitted Chris an apple.

f. 'X INTENDS TO CAUSE Y TO RECEIVE Z'
Example: Joe baked Bob a cake.

A final problem for listing approaches is that argument structure patterns change over time. Some verbs, such as *progress* and *converge*, are now being used transitively (1.11, 1.12), although they were previously intransitive (both are listed as having intransitive subcategorisation frames only in the COMLEX computational lexicon, Grisham et al., 1994). Recently there has also been an increased

[6] US Marine Spokesman Major Robert Jordan on BBC Radio report, 23rd October 1983.

[7] *The Madisonian* Newspaper, Montana, February 1889. [http://www.stcboston.org/archives/articles/nouns.shtml]

[8] Goldberg (1995, p.6)

[9] Identifying sense distinctions is rated by editors as one of the most difficult tasks in dictionary making (Kilgarriff, 1997a, p.1), and this translates into a lack of uniformity across dictionaries. Robins (1987) states that "A principled basis for decision cannot [...] be an automatic algorithm merely requiring application".

use of nouns being used as verbs directly (1.13, 1.14), replacing light verb constructions such as *have an impact* and *make a transition*.[10]

(1.11) "There is a very detailed statutory review process required prior to the designation of any new university and there are significant wider issues that need to be considered in **progressing** that.[11]

(1.12) Scientists will be discussing what has been dubbed "Tower of Babel" technology - software that can **converge** different wireless gadgets into a single device.[12]

(1.13) The Sudanese government has said it will not agree to the African Union force in Darfur being **transitioned** into a UN force.[13]

(1.14) He added recent developments had removed "a number of near-term uncertainties that have been **impacting** the company recently".[14]

Communities of speakers can quickly innovate distinctive speech patterns, particularly when their experience of an event type differs from that in the wider community. The verb *transition* below seems to have emerged from the assimilation of returning of soldiers to civilian life in the United States at the end of the Second World War.[15]

(1.15) They were being "transitioned" into teachers of mathematics, physics and engineering. I asked if this were not a form of beating "swords into plough shares" ...[16]

The verb *hold* is used intransitively in commercial air transport – while an air traffic controller can speak (transitively) of holding an aircraft at a particular altitude, while it waits to land, pilots also speak of holding (intransitively) at that altitude. Confusion between this technical use and the word's more conventional sense has caused at least one accident.[17]

Perhaps the earliest account in the Chomskyan tradition that attempted to generate argument syntax entirely on semantic principles was Fillmore (1968). Fillmore suggested an inventory of semantic classes of argument (initially termed *cases*, later called *roles*) which determined the form of their syntactic realisation – in particular what arguments would be realised as subject and direct object and what prepositions would introduce other phrases. However this and other related approaches suffered from profound problems of decidability (Fillmore, 1977a; Dowty, 1989). It was not clear what criteria could be used to put an upper limit on the number of roles in an inventory, nor how each argument of an utterance could be uniquely and unambiguously assigned a single role label (Fillmore's and related approaches, and their limitations, are discussed at length in section 2.3).

[10] The first attested examples *impact* and *transition* as transitive verbs recorded by the Oxford English Dictionary are 1945 and 1975 respectively, and they appear to have originated in technical usage in North America. Taking newspaper archives as a guide, their use appears to have increased approximately three-fold and ten-fold respectively over the last twenty years. They are now being used productively in non-technical contexts outside of North America, for example by speakers of South African (1.13) and British English (1.14).

[11] Department of Education spokesperson, quoted in the Irish Examiner Newspaper, 11th September 2006.

[12] BBC News Website, 27th September 2006 [http://news.bbc.co.uk/2/hi/technology/5382086.stm].

[13] Jane's Information Group Website [http://www.janes.com/defence/news/jdw/jdw060911_1_n.shtml, viewed 11th September, 2006].

[14] "Smallpox scare is good news for Acambis", *Guardian* newspaper, 13th September 2006.

[15] Half the stories returned by the Google News Archive [http://news.google.com/archivesearch] before 1970 concern military themes.

[16] *Gettysburg Times* Newspaper, October 30th, 1960.

[17] Stephen Cushing, 1994, p.11: *Fatal Words: Communication Clashes and Aircraft Crashes*.

Localist approaches (Gruber, 1965; Jackendoff, 1972, 1983, 1990) attempted to describe verb-argument relations in terms of analogy to the spatial domain (section 2.4). For example, the participant *Mary* in (1.16a) would be classed as a Goal argument, because it shares the preposition *to* with the physical endpoint (Goal) *Belfast* in the clause (1.17a).

(1.16) a. Post the letter to Mary

b. Post Mary the letter

(1.17) a. Post the letter to Belfast

b. *Post Belfast the letter

While this approach suffered from similar decidability problems to those of non-localist role inventories, it contributed some innovations to the literature on linking. Jackendoff (1972) proposed that a hierarchy of semantic roles, similar to that used by Fillmore (1968) for subject and object selection, could also determine the realisation of structures such as reflexives and passives (see Levin and Rappaport Hovav, 2005, chapter 6 for a view of hierarchy based approaches). He also used a predicate decomposition of the semantics of individual verbs to describe their behaviour and the role class of their arguments. For example the verb *break* might have a lexical entry such as:

(1.18) cause(Agent,change(Theme, Not Broken, Broken))

Predicate decompositions of this type have been used by many authors (e.g. Rappaport and Levin, 1988; Pinker, 1989; Croft, 1993; Krifka, 2004; Beavers, 2004) to posit semantic representations that would account for patterns of diathesis realisation (see section 2.5). An early account by Pinker (1989) suggested that verbs that could appear in the dative diathesis (such as *bring, send, give*) had two closely related but distinct senses: one involved movement of the Theme towards its Goal, while the second entailed final possession of the Theme by the Goal. The fact that locations cannot possess things, would rule out the sentence (1.17b). Similarly, the *spray/load* diathesis has been seen as expressing holistic affectedness of its direct object (Rappaport and Levin, 1988): the *hay* is fully loaded in (1.19a), and the *truck* in in (1.19b).

(1.19) a. John loaded hay onto the truck.

b. John loaded the truck with hay.

While these are simple and elegant accounts, clear counter-examples can be found that demonstrate their rigidity. In (1.20) the double object construction is used despite the Goal *him* never being in possession of the Theme *birthday cards.* And it seems implausible in (1.21) that the amount of *pesticide* sprayed is being asserted, since it is in a generic reference.

(1.20) Grant Mann would have been 31 last Friday. **His family sent him birthday cards**, but he never received them, he was already dead.[18]

(1.21) EPA's decision would make the acceptable exposure level of the pesticide one-tenth of what's currently allowed, the Post said. Farmers will still be allowed **to spray the pesticide on crops**, but its agricultural use will be reduced to a degree not yet decided.[19]

[18] British National Corpus (BNC, Burnage and Dunlop, 1992) text K1K: Central Television news script.

[19] Associated Press story on environmental warning; reposted to a discussion forum on keeping skunks as pets, viewed August 2006 [http://www.rdwarf.com/pipermail/skunks/2000-June/003016.html].

David Dowty (1989; 1991) addressed the brittleness of predicate decomposition approaches and the decidability problems of role approaches by making use of Rosch's (1973) notion of prototypes to allow role membership to be graded. In this way an argument may be a strong ProtoAgent (in which case it will likely be realised as a subject), strong ProtoPatient (in which case it will try to appear as direct object) or it may be neither (tending to appear as an oblique). When two arguments are similarly good candidates for an argument position, then either or both may take that position. This flexibility accounts in part for truth-conditionally equivalent diatheses, such as the dative and benefactive diathesis, and for the existence of synonymous verb pairs with opposing argument structures, such as *please/like* or *fear/frighten* and *meet* or *resemble.* However Dowty cannot explain cases in which one or other of the diathesis variants is preferred for pragmatic and text structuring reasons (a full discussion is given in section 2.6).

The structure of discourse can also have effects on the realisation of argument structure. It is uncontroversial that in a wide range of languages, given information tends to appear earlier in a sentence than new information, to provide conversational continuity and aid the listener in placing the new content to come (this is termed cohesion). There is also a widespread tendency for long and unwieldy sentence constituents to be placed at the end of the sentence if possible – consider complement clefts, that relocate a subject complement after the verb (1.22b), and "postponements" (1.23b) where a relative clause dislocates from its noun phrase to move after a verbal particle (Biber et al., 1999, p.935). This may also interact with verb diatheses, such as (1.24).

(1.22) a. [That she had overcome her kleptomania and become a politician]$_{\text{comp}}$ was a source of pride to her

b. It was a source of pride to her [that she had overcome her kleptomania and become a politician]$_{\text{comp}}$

(1.23) a. ?He found something [that would change things forever]$_{\text{relt}}$ out$_{\text{part}}$[20]

b. He found something out$_{\text{part}}$ [that would change things forever]$_{\text{relt}}$

(1.24) a. I baked you$_{\text{iobj}}$[those currant buns that your friend from work that you introduced me to yesterday was talking about]$_{\text{dobj}}$[21]

b. ?I baked [those currant buns that your friend from work that you introduced me to yesterday was talking about]$_{\text{dobj}}$ for you$_{\text{iobj}}$

The perspective chosen by a speaker or writer can also have an effect on argument structure (cf. Talmy's 1985 figure/ground and the viewpoint and focus of Langacker, 1990). The difference between the verbs *buy* and *sell* can be seen as a choice of which participant (the purchaser or the vendor) is to be given prominence or has particular salience. If such alternate verb pairs are not available, the passive construction may be used to place what is ordinarily a verbal complement in the subject position. Some authors have also observed a degree of egotism in the choice of argument structures. Kuno (1976) suggests that speakers prefer to speak in the first or second person, than the third person, if available argument structures make this possible, and Fillmore (1977b) suggests that human, active participants have privileged access to the subject position. The "natural topic hierarchy" of Croft (1991) is essentially an anthropocentric scale for argument selection, by which speakers give prominence to situation entities with which they are more likely to identify or empathise.

[20]The question mark symbol '?' is used to mark sentences that are questionably wellformed. See the appendix (page 183) for a list of marks and other notations used.

[21]Variation on attested use from BNC text CK9: Fiction, *The Rag Nymph*, Catherine Cookson, 1992.

It should be clear from this discussion that the mechanisms that determine the realisation of argument structure are varied and controversial. Similar controversy continues on the nature of syntactic representations. Chomskyan models assume that syntax is entirely autonomous (Chomsky, 1986, 1995) of other parts of the linguistic faculty. Lexical Functional Grammar (Kaplan and Bresnan, 1982) allows interaction between distinct levels of representation for syntax, semantics and pragmatics, while Head Driven Phrase Structure Grammar (Pollard and Sag, 1994) adopts a monostratal approach. On the other hand, experimental work, such as that by Tanenhaus et al. (1995) and Hagoort (2003) has suggested that both extralinguistic context, and general background knowledge, are quickly and apparently seamlessly integrated into online language processing (this evidence is discussed at greater length in section 4.1). Views of modularity (e.g. Fodor, 1983), have also faced theoretical challenge (Vigliocco and Hartsuiker, 2002).

As a result of these fundamental disagreements about the nature of the linguistic faculty in general, and about argument structure linking in particular, an empirical approach is taken in this thesis. First of all, as few assumptions as possible are made (these minimal assumptions about linguistic representations are described in the following section 1.1). An attempt is then made to identify what linguistic features have the strongest effect on the process of linking, or association with realised argument positions (chapter 3). In particular, features of verbal arguments, and of verb-argument relationships are examined – in common with some authors (e.g. Fillmore, 1968; Dowty, 1991) and in contrast with others (e.g. Levin and Rappaport Hovav, 2005).

The remainder of this chapter deals with some preliminary assumptions: on semantic and syntactic categories (section 1.1); on equivalence of meaning between paraphrases, and so diathesis variants (1.2); and on the nature of formal grammaticality and judged acceptability (1.3). An outline of the thesis follows (1.4).

1.1 Semantic and Syntactic Assumptions

In this thesis minimum assumptions are made on the nature of syntax. One syntactic quantity that is assumed is grammatical function. Nominative languages generally have syntactic means for differentiating subject (usually identified with causers, and with topical prominence) and object (identified with affected parties) from each-other, and from more oblique arguments (Keenan, 1976; Dryer, 2005). Many also have a fourth function, that can be termed the indirect object, which patterns with marginally affected but animate participants (Palmer, 1994).

In the context of diatheses, canonical grammatical function is referred to. It is to be understood as the grammatical function that an argument would assume in a neutral unmarked case. If a speaker is to understand the nature of participation encoded by a verb's relationship with its arguments, canonical grammatical function must be specified in the grammar: for example to recognise that the participant that is the subject of *buy* may also appear as the *from*-phrase of *sell*, or to recognise the semantic parallelisms among the variants in (1.25).[22] The alternative terms "logical subject/object" or "deep subject/object" are avoided, out of a desire to avoid making commitments to a linguistic representation of utterances that differs from surface structures. This is roughly equivalent to a Davidsonian representation (1.25e), where predicate argument positions are conventionally linked to the English verb with which it shares its name, and grammatical function is assumed from the

[22] Such linking patterns may be specified in the lexical entry of every verb, or they may only be specified for verbs whose patterns diverge from general linking principles, such as *undergo* and *receive* (which take a patient-like participant as subject), or pairs such as *be afraid of/scare* or *like/please.*

ordering (conventionally Subject–DirectObject–IndirectObject–Other). However this does not entail that a fully articulated predicate representation of utterances has an independent existence. Such structures are only assumed in the lexicon. Such a semantic representation are not implausible, but empirical evidence is lacking to settle on its precise form.

(1.25) a. John sold a dud to Mary

b. John sold Mary a dud

c. A dud was sold by John to Mary

d. Mary was sold a dud by John

e. sell(e, John, Present, Mary) $\wedge$ vendor(e, John) $\wedge$ goods(e, Present) $\wedge$ buyer(e, Mary)

For similar reasons, no appeal is made to the argument/adjunct distinction in this thesis. The problems with this division are well known – for example, attempts to annotate argumenthood and adjuncthood were abandoned in the Penn Treebank project, as consistent guidelines for identifying them could not be arrived at (Marcus et al., 1994).

This is the case particularly in the divergence between semantic and syntactic demands for arguments. Syntactic adjuncthood is typically defined by optionality – that is whether it can be omitted or not. Semantic adjuncthood is defined by compositionality – that is the contribution it makes to the description is independent of the meaning of the verb to which it applies. For example the verb *eat* could be said to have Food as a semantic argument, but it may go syntactically unrealised. Similarly a Location and Time argument is semantically necessary for every concrete verb (as tangible things necessarily happen sometime and somewhere), but these are syntactically optional for most verbs, though required for some (e.g. *originate*, *last*).

However from the point of view of the work described here, all relationships between verbs and their dependent clauses have semantic content, and therefore deserve to be described, as wide coverage role systems such as UNL do (Uchida and Meiying, 2001 – see next chapter). For the purposes of this thesis, all dependent clauses of a verb will be termed arguments.

On the referential nature of the semantic and discourse features considered here, a cognitive intensional model is followed. When deciding whether such and such an entity is, for example, Affected, it is the proposition encoded in the utterance that is referred to. Features should not be seen as attributes of verbs or nominals. That is, the subject of an agentive verb is not necessarily volitional, even though it may generally be the case (one can break something accidentally, for example) and an animate nominal is not necessarily aware (consider a person being pick-pocketed, or a patient in a coma). The proposition encoded in the utterance is considered independently of any modal or attitudinal predicates under which it is embedded – that is independently of both its ultimate truth and the degree of confidence the speaker is expressing in it.

If this approach is correct, one can make predictions about argument structure realisation on the basis of propositions, their truth values, and the participants involved. Consider the proposal that the availability of the passive construction in English is increased in cases in which the canonical object is Affected (Jackendoff, 1972 makes such a proposal). For example, in (1.26a) *him* can be viewed as Affected by the impact of the rubber bullet to his face, both physically and psychologically, and the passive (1.26b) is wellformed. If such features were modulated by the ultimate or believed truth of the utterance, expressing doubt or negating the proposition (and so reducing the Affectedness of *him*) might be expected to have consequences for the acceptability of the passive construction. However (1.26c-e) all seem as wellformed as (1.26b).

(1.26) a. A rubber bullet hit him on the face

b. He was/got hit on the face by/with a rubber bullet

c. He might have been/got hit on the face by/with a rubber bullet

d. I think/suspect/doubt that he was/got hit on the face by/with a rubber bullet

e. He wasn't/didn't get hit on the face by/with a rubber bullet

If on the other hand the feature of Affectedness was specified in a lexical entry as Dowty (1991) suggests (and it is reasonable to assume that the entry for the canonical object of the verb *hit* would involve affectedness), one would expect the passive construction in (1.27b) to be similarly well-formed to (1.26b-e). The fact that it is not suggests that the contribution of semantics to linking may be independent of modality and truth-conditions.

(1.27) a. A snowflake hit him on the face

b. #He was/got hit on the face by/with a snowflake[23]

1.2 Equivalence of Meaning

This thesis deals with meaning-preserving diatheses, and alternative choices of lexicalisation, whether within or between languages, and so presupposes the notion of sentential synonymy. While the importance of differences in pragmatic weight, or discourse function is recognised and treated extensively, the notion of a relatively fixed "literal" or "propositional" meaning underlying sentence variants is also appealed to.

Quine refers to this as "empirical meaning" (1964, p.460), which is "what remains when, given discourse together with all its stimulatory conditions, we peel away the verbiage." In reference to truth equivalence between sentences, he is interested in "broad synonymy [where] two sentences command assent concomitantly and dissent concomitantly" (Quine, 1960, p.62). Frege seems to agree that certain diatheses do not change meaning:

> The sentences 'M gave document A to N', 'Document A was given to N by M', 'N received document A from M' all express exactly the same thought. (Frege, 1897, p.141)

The same notion is appealed to in the theory of translation. For example in legal translation, the objective is the "fidelity" of the target language text, and a differentiation is made between the "situational equivalence" (i.e. literal meaning) and "pragmatic correspondence" (i.e. similarity of form) (Šarčevic, 1997, pp.2, 16, 79). In general translation, Larson (1984, pp.225) bases her Meaning-Based Translation approach on "primary meaning" which may be unchanged between several forms "which encode the same proposition". Though Barnwell (1974, pp.170-171) stresses that "focus, emphasis and prominence" cannot be ignored in translation, she also recognises a "referential meaning" that remains unchanged by passive and dative diatheses. Stylistic or pragmatic factors may drive the choice of diatheses: Baker (1992, pp.99-110) mentions the use of the passive in technical English to convey objectivity, and Larson (1984, pp.245-247) comments on its use in Hebrew as a device to avoid direct reference to the Jewish god, but neither denote a propositional meaning that differs from their active counterparts.

[23]The hash symbol '#' is used to mark sentences that are semantically illformed. See the appendix (page 183) for a list of marks and other notations used.

The relative stability of meaning through diatheses also has some experimental support. Honeck (1973) describes an experiment in which participants were asked to judge similarity of meaning between sentence variants. Base active sentences (such as 1.28a) were compared to several variants: passive versions (1.28b); active versions with nouns and verbs replaced by near-synonyms (1.28c); passive versions with near-synonyms (1.28d); and others in which canonical grammatical functions were exchanged (1.28e). Participants judged that the passive diathesis changed meaning least, while exchanging canonical functions changed meaning most.

(1.28) a. The struggle evoked the feelings that changed the lad

b. The feelings that changed the lad were evoked by the struggle

c. The fight produced the emotions that altered the boy

d. The feelings that changed the lad were evoked by the struggle

e. The feelings evoked the struggle that changed the lad

While some authors have claimed that the dative diathesis encodes subtle differences in semantics, attested examples such as (1.20) and (1.21) suggest that this may not be the case. In the diathesis experiments described in this thesis, acceptability judgements are heavily contextualised (with surrounding authentic text, and background descriptions) to constrain divergent interpretations of diathesis variants.

1.3 Grammaticality and Acceptability

Opinion differs on the nature of grammaticality or acceptability. An early scholar of Arabic, Sībawayhi (see section 2.2.3) distinguished utterances as being 'straight' (intelligible) or 'crooked' (those from which no meaning could be extracted). 'Straight' utterances were further divided into those that were 'good' (wellformed) and 'bad' (illformed but still meaningful). Additional categories of acceptability that he appeals to are felicity (infelicitous sentences are 'weak') and frequency – something might be attested in use, but (in the view of a prescriptive grammarian) illformed (Bohas et al., 1990, pp.40-41).

It is common practice among contemporary linguists to use the asterisk to indicate that (in their, or their colleagues' opinion) an utterance is not grammatical, and the question mark for those that are marginally grammatical. The question mark may also be used to signal a lower degree of confidence in the grammatical status of an utterance, and linguists often resort to ad-hoc combinations such as "*?" and "**" to express finer distinctions, though their precise usage varies (Schütze, 1996).

In Chomskyan grammars, wellformedness is entirely a syntactic phenomenon (Chomsky, 1957, p.15), independent of intepretability or plausibility: such a grammar would admit the first sentence below while rejecting the second, on the basis that the sequence of categories in the first (Adj-Adj-N-V-Adv) is licensed.

(1.29) a. Colorless green ideas sleep furiously

b. Furiously sleep ideas green colorless

While Chomsky has at times suggested that there may be degrees of grammaticality, any semantic factors that pattern with it are, in his view, reflexes of syntactic phenomena (Chomsky, 1964). For example in the case of the sentences below he sees the marginal acceptability of (1.30b) as resulting

from selectional restrictions being violated. These restrictions and the semantic properties on which they depend (such as animacy) are specified as syntactic features in the lexicon, so an utterance can be licensed without recourse to interpretation. The more fundamental phrase-structure rule violations of (1.30c) result in more profound ungrammaticality for Chomsky. Chomsky later modified this view by appealing to a competence/performance dichotomy, seeing grammaticality as a feature of (the inaccessible) competence and acceptability a (measurable) feature of performance (see Schütze, 1996, p.20).[24]

(1.30) a. John plays golf

b. Golf plays John

c. John plays and

Optimality Theory is an elaboration of Chomskyan principles in which the wellformedness of an utterance depends on the full set of candidate utterances that could communicate the same meaning or fulfil the same communicative goal (Prince and Smolensky, 1997). A strict hierarchy of discrete and universal constraints is used to select the *optimal* and sole grammatical candidate utterance, discarding all other candidates as ungrammatical. It assumes that weaker constraints cannot 'gang up' on stronger ones, meaning that relative grades of ungrammaticality cannot be expressed (Legendre et al., 2006). Variations on Optimality Theory that permit ganging-up of constraint violations and allow for grades of grammaticality have also been proposed (e.g. Linear Optimality Theory, Keller, 2006).

Some other grammatical theories appeal to an absolute notion of grammaticality, such as that found in classical Chomskyan approaches, but admit considerations beyond syntax. In both Lexical Functional Grammar (Kaplan and Bresnan, 1982) and Role and Reference Grammar (van Valin, unpublished) syntactic, semantic and pragmatic levels of representation are mutually constraining, and all must be wellformed and mutually consistent to yield a grammatical utterance. Head-Driven Phrase Structure Grammar (Pollard and Sag, 1994) demands unification (mutual satisfaction) between its syntactic and semantic representations to construct a wellformed representation, and so allows the power of predicate logics to contribute to grammaticality.

However, there is ample empirical evidence that grammaticality is a more complex quantity (Schütze, 1996; Sorace and Keller, 2005). Psycholinguistic research has identified several behavioural correlates of violations of linguistic wellformedness. Reaction-times in making grammaticality judgements have been used to chart the development of language in both first and second language learners, varying continuously as a function of acquired competence (Wulfeck, 1993; McDonald, 2000), and with the nature of the violation (Kail, 2004). Eye-tracking experiments can detect difficulties in parsing and comprehension during reading – eye-movements regress (backtrack) to earlier points in a sentence, and the durations of these regressions vary with the degree of lexical and structural ambiguity (Rayner, 1998, pp.375, 387, 390-391).[25] The location and duration of these regressions can be used to study the processing of syntactic phenomena such as long distance dependencies (Traxler and Pickering, 1996).

[24] This approach has come in for considerable criticism (e.g. Wasow and Arnold, 2005, p.1484), for two reasons. Firstly, it is curious to attempt to study an inaccessible phenomenon (grammaticality) when an accessible one is so easily accessed (acceptability). Secondly, any language phenomena that do not conform with a pre-existing theory can be dismissed as being subject to often unspecified "extralinguistic" performance factors. A theory made on such a basis is unfalsifiable.

[25] Eye-tracking equipment uses video cameras to identify the direction of gaze of each eye, by following the position of the pupil and cornea relative to each other over time. In conjunction with computer presentation of a written text, the direction, speed and duration of eye-movements (sacades) can be charted over the text, and the position and duration of fixed gazes (fixations) can be localised to words or parts of words.

Electroencephalography (EEG) can be used to detect variations in voltage that have been shown to be elicited by either semantic or syntactic violations (Osterhout and Nicol, 1999).[26] The N400 voltage component is seen in response to semantic anomalies, such as violations of selectional restrictions (e.g. *The cat won't bake ...*), while the P600 component appears for stimuli with low level syntactic anomalies, such as lack of agreement in inflectional forms (e.g. *The cat won't eating ...*). However later work suggests that the "syntactic" P600 component also appears in response to thematic incongruities (such as the syntactically well formed *The hearty meal was devouring ...*; Kim and Osterhout, 2005), that the "semantic" N400 component can be elicited by sentences that demand syntactic restructuring (such as garden-path sentences; see Bornkessel and Schlesewsky, 2006), and that the amplitude of an N400 is increased by a simultaneous syntactic violation (Hagoort, 2003, combined semantic and syntactic anomalies within single noun phrases of Dutch, with structures such as $\textit{The}_{\text{singular}}$ *salty* $\textit{craftsmen}_{\text{plural}}$). These complications, together with the fact that the P600 component is also elicited by "oddball" stimuli in other modalities (e.g. a single image of an artefact, presented among a long sequence of animal images) means that it cannot be viewed as a unique identifier of grammatical violations. Imaging studies have also identified areas of the brain that respond differentially to syntactic and semantic violations.[27] Ni et al. (2000) found increased activity around Broca's area for verb form anomalies (e.g. *Trees can grew*) and around Wernicke's area for semantic anomalies (e.g. *Trees can eat*). Newman et al. (2001) also elicited differential activations in both hemispheres of the brain for syntactic violations (word order errors, such as *I cut Max's with apple caution*) and semantic violations (implausible sentences such as *I sailed Todd's hotel to China*). However Kuperberg et al. (2000) failed to find any significant difference in activation between violations of verb selectional restrictions (e.g. *My mother ironed a kiss*) and verb syntactic valency (e.g. *His father chattered the umbrella*), despite having detected differences for wellformed but less plausible sentences such as *The woman painted the insect*.

It has also been suggested that perceived acceptability, particularly markedness, is a function of frequency – that is we find an anomalous sentence strange, simply because we rarely or never hear such a structure. Even nominally ungrammatical phrases such as *speak proper* can become acceptable, either by habituation, or a deeper process of grammaticalisation (see e.g. Tabor, 1993). For example, Greenbaum (1977) showed a very high correlation between judgements of the frequency and acceptability of several constructions (including the passive, subjunctive, and past/perfect tenses in English), and others have seen the statistical probability of a particular utterance, either as a string of words or in terms of its likely syntactic structure, as being a possible operationalisation of grammaticality (see e.g. Manning and Schütze, 1999).

Acceptability, claimed usage and actual usage do not always correspond to each other neatly (Schütze, 1996, p.60). For example, Murphy and Vogel (2007) demonstrate that while the observed corpus prevalence of adjective co-ordinations (e.g. "black and white" versus "white and black")

[26] Electroencephalography most commonly records small variations in voltage at the scalp, which are induced by neural activity within the skull (recordings under the skin of the scalp, or within the skull are also possible). The most common paradigm for analysis are event-related potentials (ERPs): small deflections of voltage that can be discovered by taking grand averages over multiple participants and multiple stimuli, of electrical activity immediately following the anomalous point in each stimulus. Various components have been identified and they are usually named for their polarity, and the latency after stimulus onset at which they peak. For example, the N400 is a negative deflection, relative to the activity seen for wellformed stimuli, that typically peaks 400 milliseconds after onset.

[27] Magnetic resonance imaging (MRI) uses a combination of a magnetic field and radio energy to excite hydrogen nuclei in the brain. Different kinds of tissue respond to this excitation in different ways, allowing anatomical images to be obtained. The level of oxygenation in blood in the brain varies with neural activity, and these differences also affect the response to excitation (this application is termed Functional MRI). In this way whole-brain images of neural activity can be obtained. fMRI has high spatial resolution relative to EEG, typically distinguishing units of neural tissue down to a scale of 2-3mm. However its temporal resolution is much lower, on the scale of seconds.

is a good indicator of judged acceptability, the same cannot be said for the English relativisers *that/which/who*: despite its being judged acceptable by participants, a sentence such as (1.31), using the relativiser *that* to modify a noun phrase that describes a human, is radically less frequent than a corresponding sentence using *who* as the relativiser.[28]

(1.31) ... Perry Mason, a fictional TV lawyer that ~~who~~ helped the poor ...[29]

Given the difficulties in using the sorts of evidence just described, acceptability judgement experiments are perhaps the most appropriate general purpose methodology for investigating the well-formedness of individual utterances, as long as certain methodological precautions are taken (as are described in sections 3.1.2.2 and 3.1.2.3).

1.4 Document Outline

Chapter 2 is an extensive discussion of varied approaches to argument structure and linking, from pre-modern times, to the Generative revolution and beyond. The reader may choose to focus on the sections on lexical decomposition (2.5), proto-type approaches, most prominently that of Dowty (1991) (2.6), and explanations based on elements of pragmatics and discourse structuring (2.7), as these are more recent, and most relevant to the work presented here.

In the empirical chapter (3), the first set of experiments (3.2) compared the relative acceptability of diathesis variants in judgement exercises with native-speaker participants of three languages: English, German and Standard Chinese.[30] Diatheses are interesting, because they constitute minimal pairs, whose interpretation and pragmatic context is fixed (contextual descriptions were supplied for all items to control interpretation). By comparing the relative acceptability of diathesis pairs, the effect of syntactic, semantic and discourse constraints can be examined, independently of confounding factors such as the overall plausibility of the sentence, the familiarity of its vocabulary, and the elegance or clumsiness of its expression. The diatheses chosen for these experiments were the dative/benefactive constructions in English, the passive construction in English and German, and the Chinese *bei* and *ba* constructions (both of which have functional and structural similarities to the passive). These constructions were chosen because they are reasonably frequent in everyday language, involve manipulations of argument structure that involve nuclear grammatical roles (subjects and objects) and because they are all well-studied in the literature. The results showed that a range of pragmatic and semantic factors have a significant effect on the perceived wellformedness of diathesis forms, some of them in more than one language. The magnitude of some effects was also comparable to that elicited by clear violations of phrase structure.

A follow-on corpus experiment (section 3.3) attempted to validate some of the results of the judgement experiments on a larger scale, by searching for associations between argument structure positions, and lexical or morphosyntactic features that are known to pattern together with semantic and discourse features of interest (for example, lexical animacy is often a prerequisite for agency,

[28] A Google search yielded 58,300 documents for the phrase "lawyer who helped" and 339 documents for "lawyer that helped". Nevertheless, in an acceptability task (Murphy and Vogel, 2007), judges gave the sentence containing "the lawyer that" a mean acceptability score of 1.3 on a 7 point Likert scale, where +3 signified "perfect", 0 signified "strange" and -3 signified "unacceptable".

[29] BNC text B21: Non-fiction, Stephanie Spindler 1991, *Learn to live*.

[30] Standard Chinese, also known as Mandarin Chinese, is the official language of both the People's Republic of China, where it is referred to as *putonghua* 普通话 ('common language') or informally *hanyu* 汉语 ('(ethnically) Chinese language'), and the Republic of China (Taiwan), where it is called *guoyu* 国语 ('national language').

and definiteness is associated with givenness). While this did provide qualified support, the effect sizes were small.

An annotation exercise among trained linguists next aimed to evaluate the level of agreement that exists between experts on the proper semantic classification of verbal arguments (section 3.4). A set of semantic role labels (also termed thematic roles) was given to classify the arguments of an arbitrary set of English sentences. The degree of agreement seen varied from role to role, suggesting that some role categories might be better defined than others. Projections for several macro-role sets were compared to estimate what level of agreement they would be expected to garner, and the results suggested that schemes such as those of Dowty (1991) and Primus (1999) are superior to traditional role sets of greater granularity.

In the conclusion, the implications of the experimental results are discussed. They support the suggestion that many types of information come to bear, and that some form of competition is at play (compare La Heij, 1988; Dowty, 1991; Prince and Smolensky, 1997) during the linking process (see sections 4.1 and 4.2). In particular, the lack of an integrated approach, that allows syntactic, semantic and discourse features to operate together in the choice of argument structures, appears to account for many of the weaknesses of previous approaches.

A possible instantiation of a numerical model of English argument linking is proposed in section 4.3, based on associations between sentence positions and semantic and discourse features that have been demonstrated to have an effect on the success of diatheses, and incorporating many elements of Dowty's account (1991). Though this linear model is numerically simple, and based on a small arbitrary set of English sentences, it can account for many of the argument structure phenomena discussed in the literature. It also provides qualified predictions of human judgements of acceptability. Finally, implications this work has for models of language processing are discussed (section 4.4), in particular whether such a model could provide a general account of the similarities and dissimilarities found in argument linking patterns across languages.

Chapter 2

Review of Role and Linking Theories

2.1 Introduction

The dominant approach to linking in linguistics has been to assume a set of semantic categories (termed semantic or thematic roles), describing the relationship between verb and argument, particularly the nature of the argument referent's participation in the event denoted by the verb. In this chapter the development of notions of roles is charted at length, from precursors that came before Chomskyan linguistics, early approaches broadly within a Chomskyan syntactic framework, localist and other decompositional approaches within the discipline of semantics, through to more recent approaches that make use of Rosch's notion of prototypes. A largely separate strand of research has looked to various discourse and pragmatic factors, such as information structure, topical salience and argument weight to explain linking patterns.

The remainder of this section introduces the notion of semantic (or thematic) roles, together with questions of the appropriate level of granularity that should be used in such inventories, and how they can be exploited in grammatical frameworks. A short review of related experimental evidence follows, before the section concludes with an outline of this chapter.

Role categories are generalisations over different notions of participation, most often used in theories of argument structure realisation. They describe the nature of the semantic relationship between arguments and the verbs that they are governed by, and are often assumed to be universal. Often they are limited to core arguments, but some approaches do reach further out into the sentences, such as the account of alternating *for/to* prepositional phrases in Wechsler (1991) or of nominalisations in Barker and Dowty (1993). A practical interlingua like the Universal Networking Language (Uchida and Meiying, 2001) aims to annotate all dependency relationships found in running text.

Roles are widely appealed to outside of lexical semantics. They are used as a descriptive device in linguistics generally, in frameworks as diverse as Head-Driven Phrase Structure Grammar, Role and Reference Grammar, Lexical Functional Grammar, and Chomskyan theories as well being assumed in many works of neurolinguistics, psycholinguistics and computational linguistics. This is despite a notable lack of consensus on what set of roles should be used, and precisely how they should be defined. The following is a "textbook" description of typical participant semantic roles (Saeed, 1997, chapter 6):

Agent the (typically volitional) initiator, or doer of an action

Patient the affected party, or undergoer of the action

Theme the entity whose state, movement or location is described

Experiencer the entity aware of, but not affected by the situation

Beneficiary the entity for which the action is performed

Instrument the entity with which the action is performed

Goal the location towards which an entity moves

Source the location away from which an entity moves

Circumstantial roles are sometimes also referred to, usually corresponding to the more oblique arguments in a sentence:

Cause the situation that precipitated this one

Condition situation that would precipitate this one

Purpose aim of the situation

Manner way in which a situation takes place

Measure extent (e.g. in cost, time or distance) associated with the situation

Location location at which the situation takes place

Temporal the time at which the situation takes place

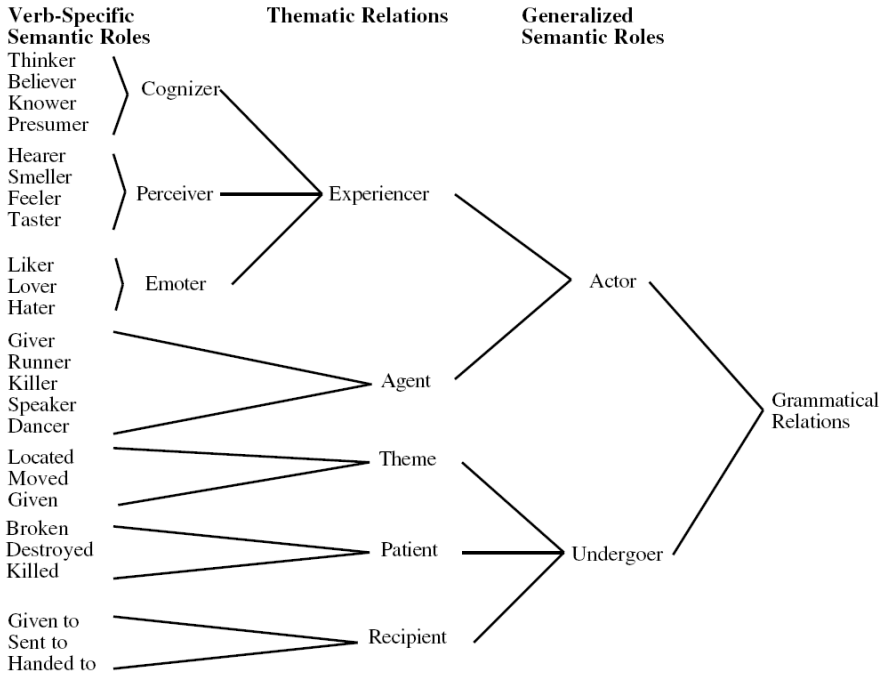
Many variations on such typologies have been suggested. For example some see the Theme as simply a marginal or neutralised Patient. Variations on Agent include Effector (or Actor, a non-sentient, but active causer) and Instigator (an Agent who orders or encourages another Agent to act, without taking part themselves). Some suggest a Recipient role, while for others that is a subtype of Beneficiary; or alternatively both are subsumed in Experiencer. Experiencers are also sometimes treated as a kind of Patient, if their resultant cognitive state can be considered a form of affectedness. The spatial Source and Goal roles are also often generalised to cover non-spatial start and end-points, describing Reasons and Consequences, or Donors and Recipients.

A related question is that of granularity. As van Valin (1999) points out, many theories share a basic two-way opposition of what he calls Generalised Semantic Roles or Macro-Roles (see figure 2.1), but disagree on how and to what extent the space below them should be carved up.[1] A compromise must be made between the detail of description provided by verb-specific roles, and the broader generalisations that can be expressed more efficiently with higher level roles. There is no obvious upper-limit on the number of roles that can be useful – Dowty (1991, p.553f) cites Lakoff who considered thirteen different varieties of Agent.

Many role based accounts specify (or assume) the uniqueness of role relationships with respect to a verb. By such a specification (e.g. Chomsky's Theta Criterion, 1981) each syntactic slot of a verb is identified with a single role in the lexicon, and no slot can share a role label with another slot of that same verb. Extending this to sentences, arguments must be attached to one and only one role, and roles must be attached to one and only one argument (this condition applies with respect to a single verb, so it is not violated by raising verbs). While this seems like a reasonable proposal,

[1] Primus (1999, 2002) proposes a third macro-role called Proto-Recipient.

Figure 2.1: Van Valin's hierarchy of role types (van Valin, 1999, figure 1, p.364)



it is not always clear which among several aspects of participation determine the role category of an argument. A Hearer may be classified as an Experiencer, but a Listener has both the causal volitional qualities of Agent, and the passive quality of an Experiencer. In a commercial transaction situation the buyer may be considered a Recipient and the vendor a Donor if one assumes that the goods are the more salient object. But since money also conventionally changes hands, the buyer can also be viewed a Donor and the seller a Recipient.

Verb diatheses, and near synonym verb pairs have also presented both puzzles and opportunities for role theories. Verbs describing situations in which there is not a clear causal asymmetry between their arguments often come in pairs, such as *like/please*, and it has often been a challenge to determine whether or to what extent their role specifications are shared, and further, how these specifications interact with the distribution of these verbs. Diatheses like dative shift or the passive pose similar challenges: for example are the roles identical between an active sentence and its passive counterpart, and if so, what is the semantic determinant for making a choice between the two forms.

In fields outside linguistics, semantic role-like categories are often appealed to. Early artificial intelligence attempts at integrating semantics into sentence generation and parsing included Wilks (1975), who used semantic clause templates such as Agent-Action-Object, together with selectional restrictions for automated analysis, while Schank and Abelson (1977) used scripts (generalised accounts of episodic knowledge) whose slots resemble roles somewhat. This idea also shows up in sociology as "recipe knowledge", that is conventionalised "pragmatic competence in routine performances" (Berger and Luckmann, 1967, p.42). Role like categories also continue to be used in machine translation systems such as Dorr (1993); Dorr et al. (2004) (which is based on Jackendoff's Lexical Conceptual Structure), Nyberg and Mitamura (2000) and Uchida and Meiying (2001).

Direct experimental support for the existence of semantic roles is somewhat lacking, as state-

of-the-art methods do not give direct access to the cognitive representations active during language processing. However a range of experiments have suggested that notions of participation and of saliency are co-active during language use, and have behavioural effects.

Early behavioural experiments in four languages (English, Italian, Turkish and Serbo-Croatian) (Slobin and Bever, 1982) established that young children (2 to 4 years old) default to a canonical sentence structure during comprehension of clauses containing unknown monotransitive verbs, interpreting subjects as agents and objects as patient. Lee (1996, p.306) found that Chinese children between the ages of 3 and 5 have difficulty producing sentences that diverge from the default clause structure, using the non-canonical *bei* (agentive marker) and *ba* (affectedness marker) constructions incorrectly.

Aphasics whose language deficit is primarily grammatical[2] have also been shown to prefer default sentence schemas of Agent–Verb–Patient (Nicol and Love, 2000) and Experiencer–Verb–Stimulus (Piñango, 2006) during comprehension and production (for a Subject–Verb–Object language). Constructions that deviate from these defaults, such as passives, or object-gapped relative clauses, cause particular difficulties for these patients (Love and Oster, 2002; Faroqi-Shah and Thompson, 2003).

Reaction time and EEG investigations of sentence acceptability suggest similar defaults. Tanenhaus et al. (1987) showed increased reaction times in a decision task on garden path sentences such as *Bill passed the test to his friend* and Bornkessel and Schlesewsky (2006) showed that ambiguous German sentences that did not conform to an Agent–Verb–Patient or Experiencer–Verb–Stimulus pattern elicited an event-related-potential (P400) that is associated with structural reanalysis.

Structural priming experiments also provide evidence that aspects of structure and meaning are active during production and comprehension. A series of experiments by Katerine Bock and others have demonstrated that the use of a particular sentence structure increases the probability of it being reproduced in subsequent utterances. For example, all other things being equal, the choice of a double-object dative construction (such as *give her the gun*) or of a NounPhrase–PrepositionalPhrase construction (such as *give the gun to her*) is biased (i.e. primed) by the presence of one or other of these constructions in preceding discourse (Chang et al., 2003). Constituent ordering alone can be primed, as a sentence such as (2.1a) can prime the superficially similar but structurally distinct (2.1b) (*ibid.* p.32).

(2.1) a. The 747 was landing by the control tower

b. The 747 was alerted by the control tower

It has also been demonstrated that both sortal features (e.g. animate/inanimate) and role features (e.g. agent/patient) of an argument operate independently in structure priming. Bock et al. (1992) found that inanimate subjects in priming sentences promoted the use of inanimate subjects in an elicited utterance that followed, regardless of structural match or mismatch (primes and elicited utterances both came in active and passive forms). And Chang et al. (2003, p.33) showed that a sentence such as (2.2) primed (2.3a) over (2.3b) (both variants are identical in terms of the grammatical categories and semantic types of their arguments, but the first is more similar to 2.2 in terms of semantic role structure).

(2.2) The man sprayed wax on the car

[2] Agrammatics, or Broca's aphasics, may be differentiated from Wernicke's aphasics, whose deficit is principally lexical.

(2.3) a. The workers scuffed dirt across the kitchen floor

b. The workers scuffed the kitchen floor with dirt

Finally, factors of discourse structuring have been shown to operate during sentence production. Arnold et al. (2000) gave pairs of participants a picture description task. The structures they produced tended to postpone heavy (i.e. long or structurally complex) constituents and those that are new to the discourse to the end of the utterance.

In this chapter the scene for semantic roles is set by examining early scholarship on argument structure, identifying ideas that were to resurface later after the advent of Chomskyan grammar (section 2.2). Then five main strands or approaches to roles for semantic description and argument linking are examined. The first (see section 2.3) is that first put forward by Fillmore (1968) of a small number of primitive atomic roles that form the core of a sub-syntactic representation to replace the deep structure of early Chomskyan theory. While the problem of proliferation of roles is a significant one, many variations on this theme have been suggested including Fillmore (1982), Somers (1987), Dixon (1991), Sowa (2000) and Uchida and Meiying (2001), some of which use situation or verb classes to constrain possibilities.

Localist approaches (section 2.4), introduced first by Gruber (1965) and developed by Jackendoff (1972, 1983, 1990), attempt to explain sentence structure via the prepositions used, and more generally by spatial analogy, though this often leads to indeterminacy of analysis. Predicate decompositional theories (Rappaport and Levin, 1988; Pinker, 1989; Croft, 1993; Krifka, 2004) instead look to verb-internal semantics to explain the argument realisation patterns of particular constructions, but their rigid specifications of what structures are licenced do not seem to account for the full range of usage (section 2.5) .

Section 2.6 describes the prototype approach introduced by Dowty (1989, 1991) which addresses the overly rigid nature of previous explanations, and which has been expanded upon for integration into both generative (Baker, 1997; Reinhart, 2002; Haiden, 2005) and constraint-based grammar formalisms (Sanfilippo, 1990; Davis, 1996).

The final strand (section 2.7) addresses the effect of discourse related factors on argument realisation. Several authors (Kuno, 1976; Fillmore, 1977b; Croft, 1991) have suggested that anthropomorphic saliency (i.e. the speaker's empathy or identification with event participants) is one determining factor in argument structure realisation, and others have added givenness and constituent weight to similar models.

The reader may choose to concentrate on the discussion of Dowty (section 2.6) and of discourse-based approaches (section 2.7) as they provide the main theoretical background to the type of model suggested by the empirical studies in the next chapter, and to examine section 2.5 as it introduces some phenomena that will be used for evaluation of a proposed model of linking presented in chapter 4.3.

2.2 A Prehistory of Linking

While Fillmore's seminal *The Case for Case* (1968) is generally acknowledged to be the first piece of scholarship to set out a theory of semantic relations to explain regularities between the syntax of sentences and their semantics, many elements of his theory were familiar to early scholarship. In this section four major cultural traditions are considered – Indic, Chinese, Arabic and European –

together with the approaches they had to explaining the relationship between verbs and their dependent clauses, in particular the manifestation of these relationships through word order, adpositions and morphological case-marking.

2.2.1 The Indic Tradition

In ancient India, the study of the sacred Hindu texts, the *Veda*, was accompanied by six auxiliary disciplines of knowledge, or *vedāṅga*. Besides *jyotíṣa* 'astronomy' and *kalpa* 'ritual', four were concerned with language: *chandas* 'meter', *śikṣā* 'phonology', *nírvacanaśāstra* 'lexicology' and *vyākaraṇa* ' grammar'. The last two, in particular, are of interest because of their significant semantic content. *Nírvacanaśāstra* was the study of words, and of how they came to bear their meaning (Kahrs, 1999, pp.xiii-xiv,48). Certain words were deemed to be *anvartha* "in accordance with that which it signifies", that is to have metaphysical inherent meaning. This is similar to the biblical account of Adam naming animals *nominibus suis* 'by their own names' (Eco, 1995, pp.7-8), and can also be compared to the irreducible core of terms in the Natural Semantic Metalanguage (Wierzbicka, 1996). More generally, a paradigmatic model of word substitution under different contexts was used to explore the semantics of individual lexemes – effectively a synonymy-based theory.

The *vyākaraṇa*, or grammatical, tradition is best known through Pāṇini's work, the *Aṣṭādhyāyī* (circa 5C BC). This is considered by some to be an early forerunner of generative grammar (cf. Lyons, 1968; Paik, 1974) by virtue of its formal and productive nature. In addition to a surface analysis of sentences by subject/predicate structure and case marking, as is seen in European traditions of grammar, Pāṇini also postulated formal categories for word-class and grammatical function, over 4,000 technical rules, and two underlying structural representations.

Central to Pāṇini's theory of representation are the *kāraka*: "factors in an action" ("*facteurs de l'action",* Rocher, 1968, p.23) which apply to "direct participants in an action" (Kahrs, 1999, p.51). Pāṇini uses them to account for the relations between semantic relations and case syntax, between passive and active sentences, and between verbs and their nominalisations. Crucially, the eight syntactic cases of Sanskrit are decoupled from the six *kāraka* categories (Verhagen, 2001, pp.278-283). Pāṇini seems to have seen the verb as central to the sentence (as seen in the next examples, the verb bears subject case), and recognised two of the cases of Sanskrit, the vocative and the genitive, as being of a different semantic kind, as they do not describe any relationship to an action. The remaining six cases (nominative, accusative, dative, instrumental, ablative and locative) are prototypically associated with a particular *kāraka*, but there is no one-to-one correspondence. Verhagen (2001, p.280) gives this translation of Pāṇini's original definitions, with modern equivalent terms:

kartṛ The [*kāraka* which functions] independent[ly with respect to other participants in a given action] is [termed] agent

karman That [*kāraka*] which the agent most wishes to reach [through the action in question] is [termed] direct object

karaṇa That [*kāraka*] which, more than any other [participant in a given action] serves as means [for the accomplishment of that action] is [termed] instrument

saṃpradāna That [*kāraka*] which one [i.e. the agent] intends as goal through the direct object [of the action in which he participates] is [termed] indirect object

apādāna The [*kāraka* which functions as a] stable point relative to a moving away is [termed] point of departure

adhikaraṇa The [*kāraka* which functions as the] substrate [relative to an action] is [termed] locus

A later, somewhat more concrete, characterisation from the *Cāndra* school (5C AD) is as follows (again, Verhagen's translation, with modern terms):

kartṛ [The *kāraka*] who performs [the action] is [termed] agent

karman That [*kāraka*] which is made [done / performed?] is [termed] direct object

karaṇa That [*kāraka*] with which [the agent] performs [the action] is [termed] instrument

saṃpradāna That [*kāraka*] to whom one wishes to give, whom one pleases, or to whom one is indebted, is [termed] indirect object

apādāna That [*kāraka*] from which one moves away, for which one fears, or from which one takes [or receives] is [termed] point of departure

adhikaraṇa The [*kāraka* which functions as the] substrate [relative to an action] is [termed] locus

Caution should be exercised when interpreting a millennia-old grammar on the basis of modern concepts, especially using limited translated materials. However, there are many striking similarities with modern Role Theories. The *kartṛ, karaṇa* and *adhikaraṇa* categories seem to resemble the roles Agent, Instrument and Location, respectively. The *Cāndra* interpretation of *karman* (identified with the accusative case) expresses the role of being created through an action, and so appears to be specialisation of Patient (other changes of state are presumably not included). Pāṇini's own description of *karman* is of an intended target, perhaps a Theme or Goal. The original definition of *saṃpradāna* (identified with the dative case) is rather vague, but seems close to the modern Goal, while the *Cāndra* broadens it to cover Recipient and Experiencer. Finally, *apādāna* (identified with the ablative case) is interesting in that for Pāṇini it is a purely spatial category (a starting point), but in *Cāndra* it has become Donor and Stimulus too, all of which could be classified as Source in a Localist system (see Section 2.4).

How the *kāraka* interact with surface morphosyntax can be demonstrated with an example. Kahrs (1999, p.52) contrasts the active and passive derivation of the sentence "the man is cutting the tree/the tree is being cut by the man". He refers to the three representations as the 'semantic level' (2.4), 'abstract syntactic level' (2.5a, 2.6a) and 'surface level of morphology' (2.5b, 2.6b). For Verhagen (2001, p.283) the same levels are 'underlying structure', '*kāraka* level' and 'case level'. The important distinction between the two underlying representations is that in the first (i.e. the deepest) the meaning is fully specified, while in the intermediate representation this has been reduced to abstracted or generalised semantic categories (the *kāraka*).

(2.4)

puruṣa	vṛkṣa	chid	lAṬ
independent-participant	desired-by-agent	verb-root	present-tense
man	tree	cut	

(2.5) a.

puruṣa	vṛkṣa	chid	lAṬ
kartṛ	*karman*		*kartari*
Agent	Direct-Object		Agentive-Present

b. puruṣo vṛkṣaṃ chinatti
man tree is cutting
Nominative Accusative Active-Present
'the man is cutting the tree'

(2.6) a. puruṣa vṛkṣa chid lAṬ
kartṛ *karman* *karmaṇi*
Agent Direct-Object Objective-Present

b. puruṣeṇa vṛkṣaś chidyate
man tree is being cut
Instrumental Nominative Passive-Present
'the tree is being cut by the man'

Both active and passive forms share the same meaning representation (2.4). Voice is chosen at the intermediate level, where it is decided which *kāraka* will be borne by the verb, and consequently, which argument will appear as the subject (compare 2.5 and 2.6). On the surface, the verb morphology indicates tense and the subject argument's *kāraka* – the nominative is "so to speak, the zero-*kāraka* case" (Verhagen, 2001, p.283). The remaining argument assumes the case most directly identified with its *kāraka* – that is, Accusative for *karman* (Direct Object) in (b) and Instrumental for *kartṛ* (Agent) in (b). A nominalisation of the event (i.e. 'the man's cutting of the tree') would also be derived from the same meaning representation (2.4).

2.2.2 The Chinese Tradition

Modern Chinese is an isolating, or uninflected language. Content words are unmarked for tense, voice, mood or grammatical functions, these functions being borne by grammatical lexemes. The rich imagery of Chinese characters, of which there are over 50 thousand (if current and obsolete variants are counted), has contributed to a strong lexical bias in linguistic studies in Imperial China. The Confucian scholar Xunzi 荀子 (3C BC), gave some of the earliest remaining reflections on language. His consideration of meaning in language was limited to the origins of lexical meaning: though "names [labels] have no intrinsic appropriateness [... or ...] reality", they were acquired through common custom, and from institutions (*ibid.* pp.143,148).[3]

It has been suggested that Old Chinese (variously defined as the language up to the end of the Western Han Dynasty, 1C AD, or Eastern Han, 3C AD) may have inflected for case (see Fuller, 1999, p.13; Pulleyblank, 1960, p.36). For example, Karlgren (1920) suggested that the opposition between the first person personal pronouns *wo* (Modern Standard Chinese: 我 'I', 'me') and *wu* (Archaic: 吾 'I', 'me') in Middle Chinese was a vestige of a nominative/objective distinction. Pulleyblank, on the other hand, believes that the two forms represent phonological variants, caused by the differing stress patterns of the canonical pre-verb position of the subject and sentence final position of the object. However there is consensus among these authors, that from Middle Chinese at the earliest (starting in the 3C AD), there was no overt marking for the core grammatical functions (i.e. subject, direct object and indirect object), and that more peripheral arguments were marked by coverbs (roughly

[3]Further, he saw it as the duty of a good ruler to regulate their use (Section 22: "Rectifying Names", p.144): "When the king sets about regulating names, if the names and the realities to which they apply are made fixed and clear, so that he can carry out the Way and communicate his intentions to others, then he may guide the people with circumspection and unify them. Hence to split words and recklessly make up new names, throwing the names that have already been established into confusion, leading the people into doubt and delusion, and causing men to argue and contend with each other is a terrible evil and should be punished in the same way that one punishes those who tamper with tallies or weights and measures. If so, then the people will not dare to think up pretexts for using strange words and throwing the established names into disorder, but will become simple and honest."

functionally equivalent to prepositions). These modern patterns of argument structure (including aspect markers, coverb usage, the form of ditransitive constructions, and the passive-like *ba* 把 and *bei* 被 structures) all stabilised during the later part of the Tang dynasty or the early part of the Song dynasty (10-11C) (Fuller, 1999, p.11; Tai and Chan, 1999, p.10).

One can speculate that the lack of a rigid system of argument structure marking in Chinese is the reason for what Peyraube (2001, p.341) describes as "... a lacuna in the Chinese [linguistic] tradition which persisted until the nineteenth century; reflections about grammar [were] practically non-existent", in particular regarding sentence and clause structure. Nakamura (1960, p.16) also comments on the lack of a tradition of formal logic, like that found in India, despite China's rationalistic tradition.

It is known that China came into contact with the Indic tradition of linguistics through the spread of Buddhism. Xuan Zang 玄奘, a Chinese Buddhist scholar who travelled to India in the seventh century gave this characterisation of Sanskrit cases (as related by his student Hui Li 慧俐 (Staal, 1972, pp.7-10)):

> "But regarding the eight inflections, the first exhibits the substance, or basis, of the thing conceived (nominative); the second exhibits the deed done (objective); the third, the means by which, and the doer (instrumental); the fourth, for whom the thing is done (dative); the fifth, what causes the thing (ablative); the sixth, whose, is the thing (genitive); the seventh, that which determines (localises) the thing, (locative); the eighth, the calling, or summoning, the thing (vocative)."[4]

It is interesting to note that the ablative is again given a derivative interpretation of reason or cause, as opposed to a purely spatial Source. The Tibetan linguistic tradition also adopted Pāṇini's *kāraka* to describe its four cases (nominative, dative-locative, genitive and instrumental; see Goldstein et al., 1991, p.32), and it is likely that this knowledge was also available to Chinese scholars.

Despite this, early grammatical description consisted only of some scattered analyses of function words and composition (Peyraube, 2001, pp.343f). Lü Yiwei (1311) 盧以緯 (original title Yuzhu 語助, later referred to as Zhuyuci 助語辭), is the earliest extant glossary dedicated to function words, in which usage is expressed by equivalence or similarity of meaning to other words (Guo Xiliang, 2003, p.51). The categories of words covered include conjunctions (e.g. *huo* 或 'or'), temporal terms (*hou* 後 'after'), adverbs (e.g. *ye* 也 'also', *zhi* 只 'only') and verbs (e.g. *yong* 用 'use'), but none of the coverbs that serve as argument structure markers. An example of the manner of definition is given below, for the adverb *mo* 莫 'not').

(2.7) "mo" you "wu" zi yi, you "wu" zi yi, you
"not" have "not/without" character meaning, have "do-not" character meaning, have
"bu-ke" er zi yi. you yi ci.
"cannot" two character meaning
"not" means the same as the word "without", as the word "do not" or as the word "cannot"
［莫］有［無］字意, 有［勿］字意, 有［不可］二字意.

2.2.3 The Arabic Tradition

The Kitāb is the earliest remaining Arabic grammar, written by the Persian scholar Sībawayhi in the 8th century. His approach was functionally, rather than formally based, and he recognised such

[4]The addition of case terms in brackets is Staal's.

modern concepts as the difference between the enunciative and the predicative (that is, between utterance and sentence), and the differing levels of acceptability and communicative goals of paraphrasal variants of a canonical sentence form (the "primary" form in his terminology) (Bohas et al., 1990, pp.38-46). Crucially Sībawayhi recognised that verb-argument relationships can remain unchanged between paraphrases.

Later, Classical scholarship took a more formal approach, in which government-dependency relations played a central role. In particular noun inflections (*i'rāb*) were viewed as marking one of several semantic values (*ma'ānī*) (Bohas et al., 1990, pp.50-60):

> The general idea expressed by many [Classical] grammarians is that the verb is a semantically complex element, composed of several internal functions, and that every one of these can be specified by an element in the sentence. In order for there to be a verb or, more precisely, a 'doing', there must be a 'doer' (fā'il), that is a subject, and a 'thing done' (maf'ūl), either as the process itself ('to hit' is 'to do a hitting'), which correspond [*sic*] to the internal object (maf'ūl muṭlaq) or as the object or objects affected by the process (one generally hits something or somebody), which corresponds to the external object (maf'ūl bi-hi), or both. Moreover, a doing necessarily happens in a specific time and place, to which correspond the 'circumstances' (ẓurūf). Lastly, to these intrinsic functions are added some others which have a less direct relationship to the process, such as the cause or motivation (maf'ūl la-hu), or the situation (ḥāl) in which the subject or the (external) object can be concomitants of the process. (Bohas et al., 1990, pp.66-67)

The passive construction challenged this view because participants other than the *fā'il* 'doer' appear in the subject position. Some scholars resolved this by stating that *fā'il* is a purely formal category, while others identified a common semantic basis among the elements that could appear in this position – that of indispensability or logical necessity, (essentially the argument/adjunct distinction: arguments could be *fā'il*, adjuncts could not) (Bohas et al., 1990, p.68).

Owens (1984, p.32-35) presents a richer typology of verbal complements, from Zamaxsharii and other classical scholars. A direct object (2.8) is one that "experiences the action of an agent [*fā'il*]". Other non-subject dependants were classified Reason (2.9), Accompaniment (2.10), Absolute (2.11) and Circumstantial, both in time (2.12a) and space (2.12b, c). As well as nominals, verbal phrases can also be governed by a main verb (2.8b, 2.9).

(2.8) a. I saw **Zayd**

b. I learned **that Zayd was leaving**

c. I passed **Zayd**

(2.9) I came to you **seeking your aid**

(2.10) I followed **the Nile**

(2.11) I really hit **him**

(2.12) a. I saw him **today**

b. I saw him **behind Zayd**

c. I saw him **at home**

2.2.4 The European Tradition

One thread common to Aristotle's *Metaphysics* and *Physics* is the origin and nature of cause and effect. In *Metaphysics* the "prime mover" or "first cause" is divine. In fact Aristotle's main argument for the existence of a god is based on this: "there must be something that originates motion, and this something must itself be unmoved, and must be eternal" (Russell, 1946, p.180). Beyond this, there were four "causes": material, formal, efficient and final. Only the third corresponds to the modern notion of cause ("efficient" having the sense of "effective"). Russell (1946, p.181) describes them by analogy: marble is a material cause, shaped on the formal cause of a statue, by the efficient cause of a chisel, wielded by a final cause, the sculptor's design. In his *Physics* Aristotle is also concerned with cause and effect, and begins a chapter with the statement that "nature (φυσις, 'physics') is the principle of movement and change" (Aristotle, 4C BC, 200b). He notes the parallelism between movement (κινησις) and change (*μεταβο*λις), anticipating later Localist models (see section 2.4). Both involve an active participant (πι*ο*ητικ*ο*ν 'doer; creator; causer') and one that is acted upon (παθητικ*ο*ν, 'undergoer'), and these correspond quite closely to the modern notions of Agent and Patient or Theme. He also recognises that there can be intermediate causes, or instruments (Aristotle, 4C BC, 256a).

Early Greek treatments of their language's case system (by the Stoics, 4C BC and Apollonius Dyscolus, 2C AD), emphasised the nature of participation with the verb (though the genitive and vocative were difficult to classify in these terms), while later work in the Greek and Byzantine traditions took up Localist ideas again (Hjelmslev, 1935, p.28-32). For example, for Maximus Planudes (13C AD) case represented both "concrete" (spatial) and "abstract" direction, with the accusative being associated with coming closer, the dative with being at rest, and the genitive with moving apart.

The terminology of roles did not appear in linguistic contexts until the 20th century, but they made their way into the English language during the Middle Ages: the Oxford English Dictionary lists the first attested examples of the appropriate sense of the term *instrument* in 1340 (2.13), *patient* in 1425 (2.14), *agent* in 1579 (2.15) and *theme* in 1620 (2.16).

(2.13) Bot þat fire..es An **instrument** of Goddes ryghtwysnes.

(2.14) Cause particler, **pacient**, or suffryng [L. Causa particularis patiens], was disposicioun of þe body ...

(2.15) The gallowes is no **agent** or doer in those good thinges.

(2.16) The externall is euery **Theme**, or matter propounded, whereof a man discourseth, or may discourse by his reason.

The earliest grammar in a European vernacular language was Ælfrid's 992 AD teaching grammar of Latin and Old English, a translation and adaptation of the 6th century Latin grammar by the Byzantine Priscian (Porter, 2002, p.1,9,31-33). Based on the classical Greco-Roman model, the text takes a peculiarly literal interpretation of case categories. The nominative is that of "naming", exemplified by "This one is named Homer, this one Vergil"; the accusative is described as being the case of "pleading" ("I accuse this person", "I plead a case"), and "more often refers to enemies", in contrast to the dative which "rather suitably refers to friends" (Porter, 2002, p.157,159).

Lily and Colet (1549, p.8) is another early teaching grammar of Latin, written in English, that describes cases in terms of the questions they answer:

> The Nominative case commeth before the verbe, and answereth to this quæstion, Who or What: as *Magister docet*, The Maister teacheth.
>
> The Genitive case is knowen by this token Of, and answereth to this quæstion, Whose or Wherof: as *Doctrina Magistri*, The learning of the maister.
>
> The Dative case is knowen by this token To, and answereth to this quæstion, To whon, or To what: as *Do librum magistro*, I give a boke to the maister.
>
> Th'Accusative case foloweth the verbe, and answereth to this quæstoin, Whom or what: as *Amo magistrum*, I love the maister.
>
> The Vocative case is knowen by calling or speaking to: as *O magister*, O maister.
>
> Th'Ablative case is communly joined with præpositions serving to th'ablative case: as *De magistro*, Of the maister, *Coram magistro*, Before the maister.

Throughout the Middle Ages, Renaissance and Enlightenment, the European grammatical tradition continued to follow Greco-Roman models. As Jespersen (1924, p.47) comments, it was "common belief that grammar was but applied logic, ... [and] a delusion that Latin grammar was the perfect model of logical consistency". An early philological description of Sanskrit makes the (assumed) correspondence between Greek, Latin and other languages explicit (Schlegel, 1808, pp.338-39):

> "Hauptsächlich besteht aber der Unterschied doch darin, daß die indische Grammatick in derselben Art, wie die griechische und römische, noch regelmäßiger, demsselben Gesetz der Struktur, wenn ich so sagen darf, noch treuer und eben dadurch zugleich einfacher und kunstreicher ist als diese".
>
> *The difference mainly lies in the fact that Indic grammar, while like the Greek and Roman and with the same structural laws, is more regular, and if I may say so, more faithful and so at the same time simpler and more artful than them.*

The tension continued between the "concrete" (i.e. spatial) interpretation of case, and its "abstract" (i.e. grammatical) interpretation (Anderson, 2004, pp.5-6). For example the dative case, corresponding to the English preposition *to*, may fundamentally be an expression of movement towards something ("concrete") and only secondarily an expression of recipienthood ("abstract"); or vice-versa. Only with Modernity do the first semantically or functionally based accounts of case systems appear. Whitney (1889, pp.88-103) lists the cases of Sanskrit (nominative, accusative, instrumental, dative, ablative, genitive, locative) by forms of usage, while Goodwin (1894, pp.xiv-xvi) has functional and syntactic definitions of Greek cases, e.g. "Accusative of Direct Object", "Accusative of Extent of Time or Space." or "Genitive of Agent or Instrument (Poetic)", "Dative of Cause, Manner, Means and Respect".

Edward Sapir was perhaps the first to apply the functional approach comprehensively across several languages. In early work he notes that the subject (or nominative) corresponds to the "doer" or "agentive, instrumental" and the object (or accusative) to "done to" or "person affected" (Sapir, 1921, pp.102, 109). Later, the International Auxiliary Language Association commissioned cross-lingual research to establish a principled basis for an artificial language intended "to serve as a sort of logical touchstone to all national languages and as the standard medium of translation" (Sapir, 1931, p.113). Sapir, Whorf and Swadesh all worked on the semantics of verbs, including aspect, and the function of argument structure positions (Swadesh, 1972, p.235).

Sapir et al. (1932) is a study of the expression of end-points in English, French and German, as part of an investigation into "Fundamental Relational Notions and their Linguistic Expression" (*ibid.*

p.3). The focus is fairly narrow, in that they do not treat derived functions of ending points such as their possible "identity [with the] 'transitive' and 'dative' relations ... nor ... kinship to goal, purpose result, etc." and "expression in the realm of time [is] scarcely touched upon." (*ibid.* p.4). The methodology is primarily to group by form, and so emphasis is given to surface phenomena that are not common to all three languages: for example pronouns with embedded end-points (e.g. *thither*); compound location-direction prepositions like *into*; implicit direction in prepositions of location (*put in the box*); and bare nouns that behave like prepositional phrases (English *home*, *aboard*; German *Heim*).

2.2.5 Conclusion

Given the importance of Greek and Latin grammar in European scholarship, descriptions of argument structure concentrated on the details of case systems. Some cases have a double personality – for example the dative case has the spatial function of denoting an end-point, and the causal function of denoting a psychologically involved participant. Presaging the attempt at a Localist description of argument structure, scholars could not agree whether both means of analysis were valid, and if so, which function was primary, and which was derived. Affectedness was recognised as an attribute of the accusative case, but causation was largely treated as a default, associated with the nominative case.

Chinese lacks any morphological case marking, so detailed descriptions of verb-argument relations are not prominent in early grammatical works, and no precursors of modern analyses can be found.

Arabic grammatical scholarship had a strong notion of dependency that runs through early and classical work. The nature of dependency relations were primarily approached formally, but some functional characterisations of core arguments, and more oblique verbal complements, were given, providing a set of semantic/pragmatic argument classes. Sībawayhi crucially recognised the invariance of verb-argument semantics across diatheses.

The work of Panini in the Indic tradition is the most interesting. His important insight was to describe semantics in a manner that was independent of syntax, providing in this way the most satisfactory account of the mismatch of syntax and semantics seen in the passive diathesis. A set of semantic primitives described the nature of participation by arguments, and provided the basis for sentence meaning. These were then realised with a 'best-fit' model of linking, in spirit not unlike the proto-roles of Dowty (1991) or the thematic hierarchy of Fillmore (1968).

2.3 Atomic Role Approaches

Charles Fillmore first proposed semantic roles. In "The case for case" (Fillmore, 1968) he proposed a variation on Transformational Grammar (Chomsky, 1957, 1958), where a representation of propositional meaning would replace the original deep syntactic structure. He viewed deep syntactic structure unfavourably as (Fillmore, 1968, p.88):

> ... an artificial intermediate level between the empirically discoverable 'semantic deep structure' and the observationally accessible surface structure, [that has] more to do with the methodological commitments of grammarians than with the nature of human languages.

Fillmore's deep case categories were semantic reflexes of surface case (which are referred to here as 'roles'), such as Agentive, Benefactive or Locative. They were proposed as "... a [finite] tenseless set of relationships involving verbs and nouns", being "semantic relationships which can hold between nouns and other portions of sentences" (*ibid.* p.2). Such a description would "comprise a set of universal, presumably innate concepts" which identify, of a given event, "who did it, who it happened to, and what got changed" (*ibid.* p.24). Individual languages would then use varying strategies at surface structure to signal these distinctions, such as word order, choice of adposition, or case morphology (realised by means of language-specific sets of transformations). The preliminary list of role categories that Fillmore proposed is listed below with their single-letter abbreviations in brackets (*ibid.* pp.24-25).

- Agentive (A), the case of the typically animate perceived instigator of the action identified by the verb
- Instrumental (I), the case of the inanimate force or object causally involved in the action or state identified by the verb.
- Dative (D), the case of the animate being affected by the state or action identified by the verb.
- Factitive (F), the case of the object or being resulting from the action or state identified by the verb, or understood as a part of the meaning of the verb.
- Locative (L), the case which identifies the location or spatial orientation of the state or action identified by the verb.
- Objective (O), the semantically most neutral case, the case of anything representable by a noun whose role in the action or state identified by the verb is identified by the semantic interpretation of the verb itself; conceivably the concept should be limited to things which are affected by the action or state identified by the verb. The term is not to be confused with the notion of direct object, nor with the name of the surface case synonymous with accusative.

With the modification 'typically animate' he includes human groupings and other anthropomorphic entities in the coverage of the Agentive role (*nation* and *robot* are the examples he gives), while the Instrumental role includes all inanimate causes. Dative corresponds to what would later be called Experiencer and Recipient, and so has a wider application than its (perhaps unfortunate) term suggests. Factitive appears to correspond roughly to a role of Patient or Product, and possibly the surface object of light verb constructions (e.g. *pay attention*) but this is not elaborated. Fillmore's only further comment on this role is that it (like the Objective role) is unusual in not being associated by default with any single preposition. One can suppose that he is trying to capture the distinction between the Objective argument *table* in (2.17) that can have things done to it, or happen to it, and the presumably Factitive argument *table* in (2.18) that cannot.

(2.17) What John did to the table was ruin it.

(2.18) ?What John did to the table was build it.

The Objective case (again, perhaps an unfortunate term, since the purpose of Case Grammar is to abstract away from surface patterns) appears to straddle the fields of Patient and Theme (themselves, not uncontroversial in the literature), and to be something of a catch-all. Further categories mentioned, but not treated as extensively, are Benefactive, Temporal and Comitative. Fillmore does say

that his list is not comprehensive and others would need to be added later. But for Fillmore, a single role for each of the locative and temporal fields is adequate, not distinguishing between Source, Goal and Path, or between Time and Duration. In his view the choice of temporal or locative preposition is determined not by semantic considerations, but rather by the head noun (consider three prepositions expressing a point in time: *on Tuesday; at 5pm; in June*), or by the verb (e.g. the verb *leave* requires no preposition on a Source argument, while *go* requires *from* or *out of*).

For Fillmore each role can only appear once in a simple sentence, other than in a co-ordinated noun phrase, and this is an important diagnostic test. For example, to establish the difference between the Agentive and Instrumental roles, he shows that while both *John* and *hammer* can appear as the subject of the verb *break* (2.20, 2.21), they cannot appear together as subject (2.23). Nor is it possible to have two Instrumental noun phrases in the same sentences, as shown by (2.24).

(2.19) The window broke.

(2.20) John broke the window

(2.21) A hammer broke the window

(2.22) John broke the window with a hammer

(2.23) *John and a hammer broke the window

(2.24) *A hammer broke the glass with a chisel

Two Agentives can be coordinated as in (2.25). If one of the two is in some way more prominent or predominant and the other is seen as assisting or accompanying (the Comitative role) the paraphrase (2.26) is also possible.

(2.25) John and Bob broke the window

(2.26) John broke the window with Bob

(2.27) *John broke the window by Bob

But this presents a problem, because *with* is usually associated with the Instrumental. If *Bob* were Agentive that would suggest the unacceptable (2.27). If *Bob* rather carries the Instrumental role, then (2.25) should not be possible, as two dissimilar roles should not be conjoined.[5]

An important notion in Fillmore's theory is that of 'case frames' (henceforth 'role frames') – that is semantic valence patterns associated with sentence types and verb classes. For example the English verb *break* seen above has the frame [O (I) (A)] (Objective, Instrumental, Agentive), signalling that it can take three arguments, in the role relationships of Objective, Instrumental and Agentive (2.22). Furthermore, the Instrumental and Agentive arguments are optional (signalled by brackets in Fillmore's notation), as either (2.20, 2.21) or both (2.19) can be omitted in surface form. Rather than considering the verb *cook* to have several senses (2.28, 2.29, 2.30), as has been suggested by Palmer (2000) and others, Fillmore sees this variation in usage as better expressed with optional frame elements: [O (A)]. This expresses the facts seen below: that the Agentive argument is optional (compare 2.28 to 2.29), and that the Objective role is either realised syntactically (2.28, 2.29), or is implicit as in (2.30), where *Mother* must be cooking something.

[5] Alternatively, it is possible (though less plausible) that *Bob* is being used by *John* in (2.26), perhaps as a battering ram, but this is inconsistent with Fillmore's characterisation of Instrumental as inanimate (*ibid.* p.81).

(2.28) Mother is cooking the potatoes

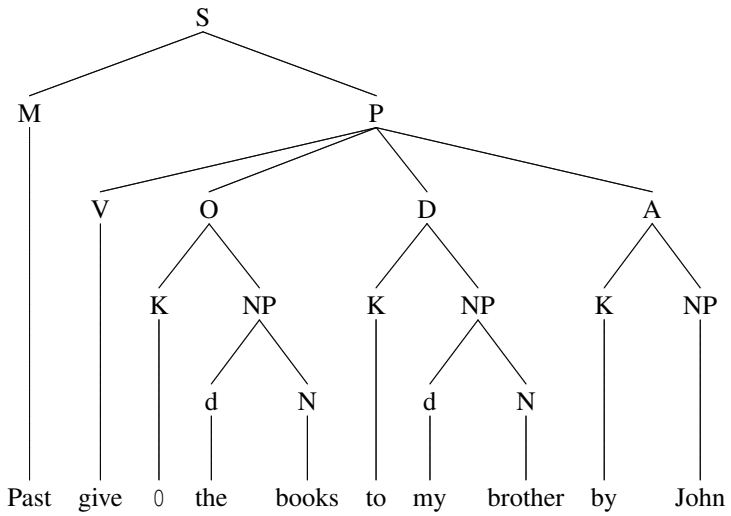
(2.29) The potatoes are cooking

(2.30) Mother is cooking

The verb *break* is also a member of the class of unaccusative verbs (such as *open, close, melt*) by virtue of having the same role frame. Which frames are admissible for a given verb is specified in the lexicon entry, as are selectional restrictions (following Katz and Fodor, 1963), syntactic subcategorisation features (including prepositions, clause type and complementisers[6]), and any linking rules particular to that verb (i.e. that depart from the general linking strategy to be introduced below). Parallelisms between semantically similar but syntactically dissimilar verbs, such as *like* and *please* are catered for by sharing the same role frame but having differing syntactic subcategorisation features – specifically, both have the frame [O D] (Objective, Dative), but *like* carries the additional information that, irregularly, D is selected as subject. Variations in frame can also account for differences between closely related verbs, such as *see* [D O] and *look* [A O], or *die* [D] and *kill* [A D]. Fillmore also claims that frames determine the generation of some sentence types, such as the imperative and the progressive, which only appear if the frame contains an Agentive argument – phenomena that earlier work by Lakoff attributed to a stative/non-stative distinction (Fillmore, 1968, p.31).

The tree below (2.31) is an example of the role categories expressed in Fillmore's deep semantic structure for the sentence *John gave the books to my brother*. Lexical categories (noun, verb, determiner) and some phrasal categories (sentence, noun phrase) are retained from Chomskyan structures[7], but the representation is order-free, and role labels, together with their associated case markers (marked K for the German *Kasus* 'case'), describe the nature of the relationship between the head verb and its dependent noun phrases. Further, the proposition P is described independently of modality (negation, tense, attitudes and the like). Fillmore does not make a model of production explicit, but he does say that the palette of frames available in a language determines what sentence types exist, and suggests that such a sentence selects a compatible verb (with some similarities to later Construction accounts, e.g. Fillmore and Kay, 1987; Goldberg, 1995), in a similar way to selectional restrictions on head nouns – an issue that will be returned to in chapter 4.

(2.31)



[6]For example, some English verbs demand *that* for sentential complements (e.g. *complain*), while others allow it to be optional (e.g. *say*, *think*).

[7]Chomskyan accounts of Case and Role are described in section 2.3.1.

At the deep level each category carries a preposition: Agentive takes *by*; Instrumental *with*; Dative *to*; Benefactive *for*; Objective and Factitive a null preposition (represented by 0 in the tree; this is *not* a trace);[8] and Locative and Temporal various prepositions, as already mentioned. The final form of a sentence is produced by moving arguments into the subject and object positions, where preposition deletion rules apply. Oblique arguments retain any preposition. The selection of an argument for subject position follows a simple hierarchy (2.32) (*ibid.* p.33): if an Agentive argument is present, it will become subject; otherwise if an Instrumental argument is present it becomes subject; otherwise an Objective becomes subject.

(2.32) Agentive > Instrumental > Objective

Moving the Agentive argument of (2.31) into the subject position, and deleting prepositions in both subject and object positions yields sentence (2.33). Alternatively the Dative argument can be promoted to the object position to yield (2.34) – Fillmore does not specify any object selection rules, so it must be assumed that this diathesis of the verb *give* is specified in the lexicon.

(2.33) John gave the books to my brother
[A ~~by~~ John] gave [O ~~0~~ the books] [D to my brother]

(2.34) John gave my brother the books
[A ~~by~~ John] gave [D ~~to~~ my brother] [O 0 the books]

The ability of a verb to passivise is also specified in the lexicon. The verb *give* can have both Objective and Dative participants in the passive subject position. A by-product of verb passivisation according to Fillmore is the loss of the ability to bear tense (hence the need to insert the auxiliary *be*), and the loss of the object-preposition deletion property. Hence the Dative preposition *to* remains in the Objective fronted passive (2.35), while in the Dative fronted passive the Objective null preposition does not need to be deleted to yield (2.36).

(2.35) The books were given to my brother by John
[O 0 the books] were given [D to my brother] [A by John]

(2.36) My brother was given the books by John
[D ~~to~~ my brother] was given [O 0 the books] [A by John]

Fillmore also explains certain locative inversions by movement and preposition deletion, treating both *be* and *have* as semantically empty verbs. A locative expression with the adjective *hot* (he believes that adjectives belong to the same semantic class as verbs) can appear in the object position, requiring an expletive *it* (2.37),[9] or can move to the subject position, losing its preposition in the process (2.38). However, he does not supply an explanation for other sentences of the form *it*–Verb–Locative that do not transform in this way (e.g. 2.39).

(2.37) It is hot in the studio
[L It]$_i$ is hot [L in the studio]$_i$

[8] This is somewhat of a puzzle, by his own admission, since the Objective and Factitive cases are associated with the preposition *of* in deverbal noun constructions, such as *The creation of the bomb* or *The repair of the city walls* (Fillmore, 1968, p.32).

[9] In fact, Fillmore describes this as involving a copying operation of the locative phrase to subject position, followed by pronominalisation to *it* or *there*.

(2.38) The studio is hot
[L ~~in~~ the studio] is hot

(2.39) a. It is raining in Sligo

b. *Sligo is raining

Similarly, the Objective phrase can occupy subject position (2.40), or in object position, taking an expletive *there* as subject (2.41). Copying the Locative expression to subject position, and pronominalising the original, results in (2.42). He does not describe how to select between these semantically empty verbs *have/be* or expletives *it/there*.

(2.40) Many toys are in the box
[O 0 many toys] are [L in the box]

(2.41) There are many toys in the box
[L There]$_i$ are [O 0 many toys] [L in the box]$_i$

(2.42) The box has many toys in it
[L ~~in~~ the box]$_i$ has [O 0 many toys] [L in it]$_i$

Diatheses of the *swarm* and *load* type (which he attributes to Partee, 1965) can be explained similarly by movement and object preposition deletion:

(2.43) a. John smeared ~~on~~ the wall with paint

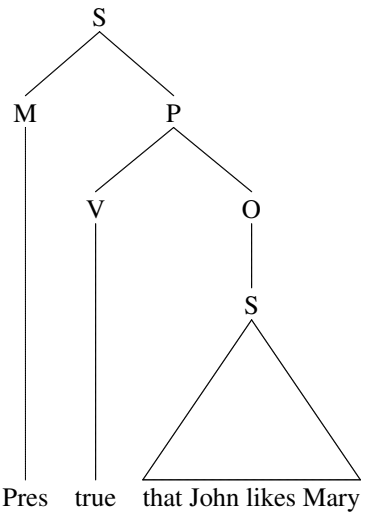
b. John smeared ~~with~~ paint on the wall

Diatheses involving clausal arguments are also covered. For Fillmore, a Sentential argument is a special case of the Objective. Sentential arguments differ in having a complementiser such as *that* or *because*, which is not subject to deletion. By similar processes to those seen in (2.37, 2.38), the sentences (2.44, 2.45) are produced from the deep structure shown in (2.46).

(2.44) It is true that John likes Mary
[O It]$_i$ is true [S that John likes Mary]$_i$

(2.45) That John likes Mary is true
[S that John likes Mary] is true

(2.46)



In a later paper Fillmore (1977a) acknowledged some criticisms of his first proposal and suggested some wide ranging changes. First of all, he conceded that it is not clear that Agentive arguments have the preposition *by* at deep structure, since non-Agentive arguments can also carry this in the passive (*destroyed by fire,* and *assumed by everybody* are the examples he gives; *ibid.* p.65). He finds the proposal that *by* is introduced by a passivisation operation just as convincing, but does not supply an explanation for the Objective and Factitive preposition *of* (suggested by its use with the arguments of deverbal nouns, e.g. *the sale of X*). He also acknowledges that there was some confusion between relational and categorial notions in his 1968 work. In particular, he sees it as a mistake to define Agentive and Dative by their animacy (*ibid.* p.66), and he sees Instrumental as a derived Patient argument that is "[set] aside ... by another [Patient] noun phrase with higher salience" (*ibid.* p.78). Rather than being driven by epistemological facts (such as animacy), his revised set of case categories should only describe relationships between arguments and verbs.

A more substantial revision is his softened position on the lack of a deep syntactic structure. Fillmore cites Anderson (1971), who showed that certain semantic generalisations held between transformationally related variants, thus supporting the existence of deep grammatical function. Anderson noted that (2.47b) is semantically anomalous, and attributes this to an association between nuclear arguments and a holistic interpretation – in this case that the subject must be affected in its entirety.

(2.47) a. Bees are swarming in the garden, but most of the garden has no bees in it.

b. #The garden is swarming with bees, but most of the garden has no bees in it.

He further claims that this restriction of interpretation (for this class of sentences, subjects and objects are interpreted holistically, while more oblique phrases can be interpreted in a partitive fashion) extends to transformed variants. Since the holistic interpretation of *bee* and partitive interpretation of *garden* in (2.47a) also apply to all the sentences in (2.48), and the interpretation of the first clause in (2.47b) holds for those in (2.49), Anderson concludes that each of these two sets of sentences share the same deep syntactic structure, and that transformations apply more plausibly after the assignment of deep grammatical function.

(2.48) a. Bees are certain to be swarming in the garden,

b. Bees were believed to be swarming in the garden.

c. John believed bees to have been swarming in the garden.

d. It is in the garden that bees are swarming.

e. It is bees that are swarming in the garden.

(2.49) a. The garden is certain to be swarming with bees.

b. The garden was believed to be swarming with bees.

c. John believed the garden to be swarming with bees.

d. It is bees that the garden is swarming with.

e. It is the garden that is swarming with bees.

Fillmore (1977a, pp.68-69) agrees with Anderson, conceding that semantic generalisations seen in the passive diathesis are also more economically stated at an independent syntactic level. For example in both variants below (2.50), *John* is an Agent and *vase* is a Patient (note that Fillmore

has adopted more conventional terminology by this time). Linking rules that related agenthood with deep subjecthood, and patienthood with deep objecthood are a more attractive alternative to differing sets of rules for active and passive sentences.

(2.50) a. John broke the vase.

b. The vase was broken by John

Finally Fillmore addresses what he sees as the most serious criticism of his work: "that nobody ... has come up with a principled way of defining the cases [roles], ... determining how many cases there are, or for determining when you are faced with two cases that happen to have something in common as opposed to one case that has two variants". He noted that linguists even have difficulty agreeing on the nature of agency (Fillmore, 1977a, p.70).

His response is that "meanings [be] relativised to scenes" (1977a, p.59). For example the distinction between various verbs associated with commercial events, such as *buy*, *sell* and *pay*, lies in the choice of which elements are to be put into perspective, that is realised as one of the nuclear arguments of subject and direct object. In Fillmore (1977b, p.95) he goes further to suggest that "we [should] ask, not what kinds of relationships there are that link [arguments] with their verbs but what properties of scenes determine whether or not something will be realised as a [nuclear argument]".[10]

2.3.1 Derivative Approaches

Fillmore's consideration of scenes led to his later work on FrameNet (1982). FrameNet categorises each of a verb's senses by the type of event it represents and the arguments that it realises syntactically. For example the *Commerce_buy* (instantiated as *buy* or *purchase* in English) and *Commerce_sell* (*sell, retail, flog*) frames are both perspectives on the *Commerce_goods-transfer* frame, and all three share the elements Buyer, Seller and Goods, among others (what Fillmore previously referred to as cases, are much more finely grained, and are termed "frame elements"). The additional detail provided by frames comes at a cost in loss of explanation. Frames record what arguments are in perspective (presumably stored in the mental lexicon), rather than determining these on the fly with general linking principles. Additionally, less generalisation across frames is achieved: the predicates *terrify* and *be afraid of* belong to unrelated frames. Both share the Experiencer element, but the source of fear differs: termed Stimulus for *terrify* (2.51) and Content for *be afraid of* (2.52).[11] He foresaw this limitation in earlier work: in such approaches "there is no concern with whether the case [of one verb class] is or is not the same as some particular case in the conceptual framework posited for some other verb class" (Fillmore, 1977a, p.72).

(2.51) Horror films$_{\text{Stimulus}}$ terrify Edna$_{\text{Experiencer}}$

(2.52) I$_{\text{Experiencer}}$ am afraid of spiders$_{\text{Content}}$

Another example of this are the verbs *destroy, break,* and *damage* whose objects are described respectively as Undergoer, Whole_Patient and Patient. With such a system it is no longer possible to state some generalisations economically, such as the fact that Patients passivise readily, or that non-Agent subjects resist the progressive form (as discussed on p.30).

[10] This quotation deals with objects, but as the preceding text makes clear, subject selection is dependent on object selection. In this later work Fillmore also considers that saliency may play a role in argument selection, as is discussed in section 2.7.

[11] Examples from FrameNet website: http://framenet.icsi.berkeley.edu/, viewed 9th August 2006.

Table 2.1: Sowa (2000) role grid

	Initiator	**Resource**	**Goal**	**Essence**
Action	Agent, Effector	Instrument	Result, Recipient	Patient, Theme
Process	Agent, Origin	Matter	Result, Recipient	Patient, Theme
Transfer	Agent, Origin	Instrument, Medium	Experiencer, Recipient	Theme
Spatial	Origin	Path	Destination	Location
Temporal	Start	Duration	Completion	PointInTime
Ambient	Origin	Instrument, Matter	Result	Theme

Many other authors have devised schemes along the same lines as Fillmore's earlier work, but usually with much larger inventories of roles. Dixon (1991) proposes thirty odd verb classes for English, with class specific roles such as Speaker, Addressee, Medium and Message (for the Speaking class) or Owner and Possession for the Own class – in effect a similar but slightly more general approach than FrameNet. Halliday's 'experiential' level of representation (differentiated from the discourse and grammatical levels) encodes similar information: material processes (2.53) have Actors and Goals (roughly, Agents and Patients), and mental processes have Sensers and Phenomena (2.54) (Halliday, 1985, pp.33-36,109-119). However, both of these proposals are predominantly descriptive, and do not provide rule-based procedures for the generation of argument structure from semantics.

(2.53) $\text{Jack}_{\text{Actor}}$ $\text{fell}_{\text{Process,material}}$ down and $\text{broke}_{\text{Process,material}}$ his $\text{crown}_{\text{Goal}}$

(2.54) But the $\text{quiet}_{\text{Phenomenon}}$ $\text{troubles}_{\text{Process,mental}}$ $\text{me}_{\text{Senser}}$ all the same

A derivative approach that has been used in the field of Artificial Intelligence is a set of related proposals by Ostler (1980), Somers (1987), Dick (1991) and Sowa (2000). They use a small number of event type classes (Active, Objective, Dative, Locative, Temporal and Ambient) and four Localist role classes (Source, Path, Goal and Local; see section 2.4). Sowa replaced these four role classes with Aristotelian terms: Initiator – initial cause; Resource – material cause; Goal – final cause and Essence – formal cause (see table 2.1). Some cells in these systems are further distinguished by atomic features. For example $\pm$Animate determines whether the Initiator of an Action type event is labelled Agent or Effector (i.e. Actor). While Sowa's event types are perhaps more intuitive than Somers', and his role types perhaps less clear, all these proposals suffer from the drawback of being able to describe at most four different verb-argument relationships. It is possible to find verbs with more than four subcategorised arguments. The verb insure can take five (2.55): the insurer, the insured property, the premium, the payout and the events that are insured against. Similarly *bet* can take five arguments (2.56): each of the two parties making the bet, the amount, the event bet on, and the odds.

(2.55) [We] will insure [your car] [against fire and theft] [to a total of 3000 euros] [for only 15 euros a month].

(2.56) [Tom] bet [me] [twenty euros] [on the final] [at two to one]

Early work by Noam Chomsky considered the possibility that verb-argument semantics play a role in clause generation. He discussed the view that subjects are the Actor of an action, and

that objects are their Goal or Object, and concluded that though "striking", such correspondences between syntax and semantics are "imperfect" (Chomsky, 1957, pp.94-101) on the basis of examples such as *John received a letter* (*John* is not the actor) and *I missed the train* (the train is not affected). He states that "meaning will be relatively useless as a basis for grammatical description",[12] and suggests that it be introduced into a "metatheory that deals with grammar and semantics and their points of connection" (Chomsky, 1957, pp.101-102). Perhaps the first comprehensive account to articulate the necessary "metatheory" was Katz and Fodor (1963), further developed in Katz and Postal (1964), which proposed that the lexicon includes semantic features that have to be unified to produce a wellformed sentence. For example the sentence *Golf plays John* would be ruled out, and *John plays Golf* allowed, because the entry for *plays* demands an animate subject and inanimate object, and the entries for *golf* and *John* would carry the features -animate and +animate respectively.

In Chomsky's theory of Principles and Parameters (Chomsky, 1981), a large part of the mechanisms of grammar were moved into the lexicon. Verbs would have both categorial (i.e. syntactic) and semantic properties that are "projected" from the lexicon to license an utterance as wellformed, or not (this is termed "c-selection" and "s-selection") (pp.60-68 Smith, 1999). Chomsky also considered that these syntactic properties might be systematically derived from semantic properties by Canonical Structural Realisation rules (CSRs). For example, the CSR of a patient would be an noun phrase, and the CSR of a proposition might be either a noun phrase or a clause (Chomsky, 1986, pp.86-87). He also discussed the possibility that notions such as Agent might be part of the "primitive [cognitive] basis", independent of Universal Grammar (Chomsky, 1981, p.10). However, this does not quite amount to a theory of linking, since grammatical role (e.g. subject or object) still has to be specified in the lexicon. The portions of his theory that dealt with the satisfaction of selection requirements projected from the lexicon were termed Theta Theory and Case Theory[13], and they have some superficial similarities to Fillmore's theories. The first is responsible for ensuring that a verb's semantic subcategorisation is saturated, and the second that its syntactic subcategorisation is saturated (compare HPSG's Content and Subcat features). Two levels of saturation were needed to account for certain "quasi-arguments" – elements such as existential *there*, expletive *it* or argument members of idioms and light verb phrases (e.g. *take advantage* or *kick the bucket*) – that are required syntactically for wellformedness, but not semantically (and hence are absent at d-structure) (Ouhalla, 1999, pp.148-152). Only noun phrases with "some sort of referential function" are full arguments, and each must have exactly one theta role (additionally every theta role must have exactly one argument – the "Theta Criterion", Chomsky, 1981, p.36). As such, Theta-Theory is essentially a mechanism for regulating the number and syntactic types of arguments – the role label itself is irrelevant, and plays no part in the semantic representation: on the basis of economy, Chomsky assumes that his semantic representation (termed Logical Form) differs little from his (previously assumed) syntactic form (an example is given in (2.57), Chomsky, 1981, p.35). In his theory-internal view, supposing instead a form like (2.58) or (2.59) would entail extra commitments. And as Smith (1987, p.73) concedes, "There has never been agreement [within generative grammar] about how many theta roles it is necessary to assume, nor what their precise semantic properties are ...".

(2.57) John_i $[_{VP}$seems $[_S$ t_i to $[_{VP}$be sad]]]

[12] It is revealing that he draws this conclusion, rather than its (equally unjustified) converse that syntax might be "relatively useless".

[13] In Chomsky's theory all syntactic arguments bear 'abstract' case, even in languages, such as Chinese, that do not exhibit it. Abstract cases are named after European cases such as Nominative, Accusative, Dative, and for our purposes here can be thought in terms of their counterpart grammatical functions of Subject, Direct Object and Indirect Object.

(2.58) seems(sad(John))

(2.59) seems(sad(e)) $\wedge$ experiencer(e, John)

In the developing Minimalist Program d-structure has been dispensed with, and so there is no longer a need for the Theta-Criterion (that every argument have exactly one role). However, Chomsky still sees the possibility that theta roles partially determine selectional and case requirements (Chomsky, 1995, pp.187-188, 236-238).

In HPSG the problem of settling on a number of roles, and their degree of generalisation is settled by giving each verb unique roles – *persuade* has a Persuader and a Persuadee, *try* has a Tryer, *give* has Giver, Givee and Gift (Pollard and Sag, 1994, pp.28-29, 134-135). While formally convenient, this degree of specificity leads to semantic vacuity. Firstly no parallels are encoded between related roles, such as Buyer and Sellee, or Giver and Donater. Secondly the role labels chosen are simply a redundant restatement using English morphology of the grammatical functions already encoded in the syntax component. For example given a sentence containing the unknown verb *clepe* (archaic 'to cry, call') "He went to the window and cleped to the people"[14], one can predict that the subject will be annotated as Cleper and the object as Clepee, without any knowledge of its semantics. Of course these labels can be more charitably seen as a description of the bundle of things that a speaker knows to hold for each argument position of each verb, but as there is no generalisation across verbs, or situation types. This approach does not add to semantic description, and the content of labels does not have any effect on clause realisation. Crucially, as the labels are lexically fixed, they cannot represent differences in role relationships of the same verb across different utterances and the interpretations they receive, such as the difference in affectedness between being hit by a snowflake and a rubber bullet (see page 9). Thus their only function is an indexing one like that of Theta Roles in Chomskyan theories, keeping track of semantic valency relations through processes such as the lexical rule for passive formation.

Perhaps the most comprehensive of these atomic role approaches are the 'relations' of the Universal Networking Language (Uchida and Meiying, 2001) – a machine readable interlingua (i.e. language independent representation of meaning) for tasks such as machine translation. Uchida and Meiying take the position that every language expresses the same space of relationships between verbs and arguments, but languages differ in how they carve it up, and so any language can serve as a practical interlingua if sufficiently disambiguated. Because of its widespread familiarity, a formalised version of English is used as the basis of UNL. The participant roles include some novelties such as Co-Agent (for independent second initiators, such as *Juliet* in *Romeo met with Juliet*) and Partner (for cooperating second initiators, such as *Clyde* in *Bonnie did all her best heists with Clyde*), but most of the differences are terminological, to make it more accessible to non-linguists: Patient is called Object, Theme is called Thing with Attribute. Since UNL is intended as a fully fledged system for annotating unrestricted running text, it includes a wide range of circumstantial role categories also, in the temporal (initial and final time; duration; sequence; co-occurrence), spatial (paths; static, initial, intermediate and final places), conditional (distinguishing prior reasons from final objectives and enabling conditions), statal (initial and final states) and other (e.g. means and manner) realms (UNL Centre, 2001, Chapter 1).

However, as is clear from the preceding text, atomic role approaches have failed to reach consensus on the correct set of roles to explain various grammatical phenomena, or even to agree on the appropriate level of granularity.

[14] Variation on attested utterance from Oxford English Dictionary.

2.4 Localist Approaches

Gruber (1965, republished as Part I of 1976) was the first scholar to articulate a localist theory of argument structure within the mainstream of the generative tradition. He proposed a level of representation called Prelexical Categorial Structure, that would underlie the deep syntactic structure of the Chomskyan theories of the day, and which he hoped would be language-universal. He adopted the semantic compositionality and selectional restrictions espoused by Katz and Fodor (1963), but his approach was 'derivational' – basically the same as that of Generative Semantics (e.g. Lakoff, 1971), with the prelexical structures forming the basis of both syntactic generation and semantic derivation.

Central to Gruber's conception of semantics is the Localist Hypothesis[15] – that is that patterns of representation in the spatial and temporal domains are used by analogy in the representation of more abstract fields. Since the prepositions *to* and *from* are used to signify start and end points in temporal and spatial terms, Gruber conceives of abstract change (that is, other than motion, or the passage of time) in similar terms: "We will consider ourselves justified in using the term 'abstract motion' or 'abstract transition' because of the similarity in the senses of what is expressed and because of the identity of the prepositions used in all these senses." At prelexical structure the representation of the motional *from* and *to* phrases in (2.60) is identical to that of the *from* and *to* phrases in sentences (2.61)-(2.64). In all these examples the entity undergoing change is viewed as the Theme (described in more detail below).

(2.60) The letter went from New York to Philadelphia

(2.61) The climate changed from being rainy to having the dryness of the desert

(2.62) The circle suddenly switched from turning clockwise to turning counterclockwise.

(2.63) John translated the letter from Russian into English

(2.64) Bill converted from a Republican to a Democrat

A sentence such as (2.65), which subsumes the meaning of (2.66), has the same prelexical structure, with the addition of the verb feature Agentive. By analogy, a purchase in which no movement takes place is treated identically (2.67).

(2.65) John sold a book to Mary

(2.66) The book went from John to Mary

(2.67) John sold a house to Mary

Information (the *that* phrases in examples 2.68, 2.69) can also be viewed as a Theme moving between abstract locations. As the previous example shows, possessors can be abstract Themes, and their possessions abstract locations.

(2.68) John learned that the earth was flat from Bill

(2.69) John communicated to me that he would not be finished on time

[15] Localism in semantics is not related to the notions of locality and long-distance dependency of syntactic theory.

Table 2.2: Some examples of Gruber's verb features (1976, p.209)

Non-Agentive	Motional	Durational
Possessional	*lose, inherit*	*keep, retain*
Positional	*fly, roll*	*keep*
Identificational	*turn, become*	*keep, remain*
Causative-Agentive	Motional	Durational
Possessional	*give, sell*	*deprive*
Positional	*make*	*keep*
Identificational	*change*	*bar*
Permissive-Agentive	Motional	Durational
Possessional	*accept*	*leave*
Positional	*drop*	*leave*
Identificational	-	*permit*

However, often a single situation type can have various prelexical representations. In Gruber's examples (2.70) and (2.71) the representation of possession varies – in the first the possession is a Location, and in the second the possessor is Location. And in the paraphrases (2.72) and (2.73) the dot and circle each are variously represented as Theme and Location. Similarly in situations that involve cognition, paraphrases expressing the same proposition can have opposing structures: the Theme in (2.74) is *I*, and in (2.75) is *sushi*.

(2.70) John is out of money

(2.71) John has the book with him

(2.72) The dot is inside of the circle

(2.73) The circle is around the dot

(2.74) I'm inclined towards sushi

(2.75) Sushi appeals to me

Unlike Fillmore, Gruber does not enumerate a set of role-like semantic categories as being drivers of argument structure realisation. Rather a general template for a sentence exists at prelexical structure, and the selection of a compatible verb imposes a particular syntactic structure, including grammatical functions. Verbs are categorised along three dimensions. Verbs must be either Motional or Durational, and must also belong to one of the types Positional, Possessional, Identificational. The third dimension is agentivity – Non-Agentive, Causative-Agentive and Permissive-Agentive. Examples of these categories are given below in Table 2.2, and as can be seen, Gruber focused on concrete spatial and temporal domains. Single features are said to explain the differences between closely related sentences: (2.76) and (2.77) differ only in the Agentive feature attached to the verb *roll*.

(2.76) John rolled the ball down the hill

(2.77) The ball rolled down the hill

While Gruber does mention categories like Agent or Goal, they are not defined independently, rather being short-hand for structural configurations (Agent is the subject of an Agentive verb, and

Goal is the To prepositional phrase of a Motional verb). Theme is the only role-like category that is independently characterised. For Gruber, a Theme argument is "semantically discernible [as] the entity which is conceived as moving or undergoing transitions" (*ibid.* p.38). A Theme argument is present in every sentence, and it is the syntactic and semantic focus, inviting comparison with the Halliday's Theme/Rheme distinction (Halliday, 1985, chapter 3), though it is easy to imagine a sentence where an unaffected entity is the semantic or syntactic focus of the sentence. In (2.78) *Ballynahinch* is unambiguously the focus of the sentence, but does not undergoes transition of any kind.

(2.78) It was in Ballynahinch that I learned to stop worrying and love the bomb

At prelexical structure, noun phrases are attached to the verb by preposition-like operators, such as At, To, Out Of or From, Through and many others. Verbs often incorporate certain elements of the situations they express, meaning that they may be left out of utterances. For example, the verb *cross* is said to compulsorily incorporate the operator Across, hence the unacceptability of (2.79a) (Gruber does not consider that this might be a simple case of lexical blocking between *cross* and *across*, as the broadly synonymous 2.79b suggests). A verb like *eat* optionally incorporates the food eaten (2.80), and the verb *send* optionally incorporates the operator To (as long as the Goal entity is of semantic type Human or Society) – hence the paraphrases (2.81a) where it is lexicalised as *to*, and (2.81b) where it is incorporated.

(2.79) a. *John crossed across the street

b. John crossed over the street

(2.80) John ate (some food)

(2.81) a. John sent a book to Mary

b. John sent Mary a book

Any operator attached to the subject is incorporated, meaning that they appear as bare noun phrases. In this way pairs of verbs such as *buy* and *sell* can share large parts of their prelexical structure – both have the buyer with the operator To, seller with the operator From (2.82). The At operator, that is a default for Theme arguments, is not lexicalised in subject position, so fronted Themes (for example in passive constructions) appear without a preposition.

(2.82) a. ~~to~~ John bought a book from Bill

b. ~~from~~ Bill sold the book to John

Gruber also develops substitution operations at pre-lexical structure to establish the truth-equivalence of paraphrases: a "semantic calculus equating or relating non-derivationally connected structures". He asserts several prelexical equalities involving negation, such as At = Not To, and Go = Not Remain and Out = Not In. Hence the identity of meanings below can be derived, as indicated in the square brackets.

(2.83) a. John is in the money [In]

b. John is not out of money [Not Out = Not (Not In) = In]

(2.84) a. John did not remain off the rug [Not Remain At Not]

b. John went onto the rug [Go To = Not Remain To = Not Remain At Not]

While Gruber's work describes many parallels between the expression of spatial notions, it does not provide an explanation for linking patterns. In particular, regularities in the choice of object and subject are not addressed, leaving this information to be enumerated verb by verb in the lexicon. And as was shown with examples (2.71-2.75), semantic equivalence between two sentences does not prevent them from having diametrically opposed prelexical structures. A second weakness, in common with many accounts of roles, is the lack of any natural limit on the number of "primitive" operators. Besides the prepositions (of which there are typically many in a single language – the COMLEX computational dictionary lists 50 spatial and temporal prepositions for English alone) he appeals to types such as To Place Intended, and Information.

Ray Jackendoff used Gruber's ideas on localism as a jumping off point, agreeing that "the semantics of motion and location provide the key to a wide range of further semantic fields". More formally, his Thematic Relations Hypothesis posits that "in any semantic field ..., the principal event-, state-, path-, and place-functions are a subset of those used for the analysis of spatial location and motion" (Jackendoff, 1983, p.188). In a series of works (Jackendoff, 1972, 1983, 1990, 1992) he developed a system of decompositional semantic representation called Lexical Conceptual Structure (LCS), based on the localist hypothesis, and involving spatially based primitive operators, such as Go, Cause, Change, To and In.

(2.85) John ran into the room
[$_{\text{Event}}$ GO([$_{\text{Thing}}$ John], [$_{\text{Path}}$ TO([$_{\text{Place}}$ IN([$_{\text{Thing}}$ Room])])])]

In Jackendoff's model Chomskyan deep structure is transformed to surface structure in a conventional fashion, but interpretation rules apply in parallel to transformations, mutually constraining each-other, and ensuring that the syntactic realisation and semantic interpretation remain compatible. There were four elements to semantic interpretation: focus and presupposition (governing intonation patterns, based on previous discourse); co-reference (anaphora relations); modal structures (negation and attitudes); and, of most interest to the present study, functional structure, that is "... relations in the sentence induced by the verbs, including such notions as agency, motion, and direction" (Jackendoff, 1972, p.3). This approach retained syntactic generality at deep structure, and avoided the elaborate feature annotations demanded by Katz and Postal's principle that all necessary information for semantic interpretation be present there.

As LCS developed (a later example is given in 2.85), role labels became less important. By Jackendoff (1990, p.47), they had been relegated to shorthand for particular representational configurations, in a similar manner to the status of the grammatical functions subject and object in Chomskyan grammars:

> "... thematic roles are nothing but particular structural configurations in conceptual structure; the names for them are just convenient mnemonics for particularly prominent configurations [of LCS]."

In the example below (Jackendoff, 1972, p.39-42), *Charlie* is considered an Agent as "the argument of Cause that is an individual", and *nut* is Theme as "the argument of Change that is an individual".

(2.86) Charlie opened a pistachio nut
$$\text{CAUSE(CHARLIE,} \begin{bmatrix} \textit{CHANGE} \\ \textit{physical state} \end{bmatrix} \text{(A PISTACHIO NUT, NOT OPEN, OPEN))}$$

However, in the first substantial presentation of Jackendoff's semantic structures (1972), role categories do have grammatical consequences, determining the generation of the passive form, and of reflexive constructions. Jackendoff follows Gruber in the characterisation of Theme as being obligatorily present and unique in every simple sentence. In spatial contexts the Theme is the entity whose location is being asserted (it is the subject in examples 2.87 and 2.88, and the object in 2.89) or that undergoes motion (the subject in 2.90 and the object in 2.91 and 2.92).

(2.87) The rock stood in the corner

(2.88) John clung to the window sill

(2.89) Herman kept the book on the shelf

(2.90) The rock moved away

(2.91) John rolled the rock from the dump to the house

(2.92) Bill forced the rock into the hole

By extension, possessions were seen as Themes also, and possessors as a Location (2.93)-(2.95), or being passed from Sources to Goals (2.96)-(2.98).

(2.93) Herman kept the book

(2.94) The book belongs to Herman

(2.95) Max owns the book

(2.96) Harry gave the book away

(2.97) Will inherited a million dollars

(2.98) Charlie bought the lamp from Max

Similarly, information was also a Theme, either at a Location (*Max* in 2.99) or moving from a Source to a Goal (from *Dave* to *students* in 2.100).

(2.99) Max knows the answer

(2.100) Dave explained the proof to his students

States and transitions in general were described by analogy: any entity to which a state is ascribed is a Theme; and Location, Source and Goal describe unchanging, initial and final states respectively. Cognitive states could be notated in these terms, cognisers being Themes (2.102, 2.104, 2.106). Jackendoff sees the ability to give a unified account of spatial and derived usages of a single verb (e.g. *keep, stay, get* and *go* in the surrounding examples) as a significant advantage of his theory.

(2.101) John stayed in the room

(2.102) John stayed angry

(2.103) George got to Philadelphia

(2.104) George got angry

(2.105) Harry went from Bloomington to Boston

(2.106) Harry went from elated to depressed

Jackendoff's use of the category Agent is more specific than Gruber's, in that it is "... identified by ... will or volition toward the action ... Hence only animate NPs can function as Agents" (Jackendoff, 1972, p.32). Accordingly he suggests the insertion of volitional adverbials, such as *deliberately* or *in order to*, as a diagnostic test for the presence of an Agent in a sentence. Similarly imperatives are only acceptable if the elided subject is an Agent.

(2.107) *The rock deliberately rolled down the hill

(2.108) John deliberately rolled down the hill

(2.109) *Receive the book from Bill[16]

Caution must be exercised with such tests for two reasons. Firstly, the acceptability of a sentence with an adverb such as "deliberately" inserted does not guarantee that the utterance (i.e. that sentence in this context) is agentive – rather it confirms that the sentence can have an agentive reading, and does not exclude the possibility of other non-agentive reading (as indeed 2.108 has). Secondly, the insertion of adverbs of volition is blocked for verbs that are explicitly volitional, due to redundancy – consider the marginal acceptability of *deliberately murder* or *try on purpose*.

Furthermore this restrictive definition of Agent presents obvious problems for inanimate causative noun phrases. In (2.110), *camper van* is inanimate, and so cannot be an Agent. Since it is moving one can assume that Jackendoff would label it the Theme. However, the *geraniums* undergo a clear transition – also a defining characteristic of Theme – but Jackendoff only allows one Theme per sentence. Another issue that is not addressed are marginal cases such as insects, that while animate, might not be quite viewed as volitional, and of artefacts such as computers or robots that are inanimate, but which may be viewed by speakers as volitional (2.111).

(2.110) The camper van crushed our geraniums

(2.111) The program wants you to press the button that says "delete all files"

Jackendoff does however allow single entities to carry more than one role label relative to a single verb. In (2.108) *John* is in one sense an Agent, as the causer of the *rolling* event, but he is also the Theme, as the entity undergoing movement. For *trading* events, the possibility of assigning multiple roles to a single entity allows him to show the commonalities between (2.112) and (2.113). In both sentences, the Theme is *pig*, Source is *Zelda*, and Goal is *Harriet*. In addition to this, money is being passed, so it is described as a Secondary Theme, with corresponding Secondary Goal (*Zelda*) and Secondary Source (*Harriet*). Finally, what distinguishes these two sentences is that the Agent is identified as "who is designated as taking initiative" – *Harriet* in (2.112), *Zelda* in (2.113).

(2.112) Harriet bought a pig from Zelda for $5.98

(2.113) Zelda sold a pig to Harriet for $5.98

[16]Exhortations such "Receive these blessings" are counter-examples, but these may be due to a difference in sense, referring to active acceptance.

While this is a neat explanation for the variation in the use of *buy*/*sell*, *get*/*receive* and other such pairs of verbs, it is easy to come up with counter-examples. In (2.114) *salesman* appears to be the unequivocal initiator of the commercial transaction, but *buy* is used instead of *sell*, probably for discourse reasons.

(2.114) A pushy salesman called to my door yesterday, and in the end I bought a Wheel-o-Vator stair lift and a 35-year subscription to the Reader's Digest

Jackendoff sees Fillmore's demand that each argument carry only one role as a fatal flaw (though this feature of Fillmore's approach is independent of the adoption of the localist hypothesis – one can easily imagine a variant of Fillmore's theory that does allow more than one case per argument). Rather than develop a large number of more finely grained role categories (e.g. Donor, or Agentive Theme), Jackendoff (1990, p.95f,125f) adds a small number of new roles (including Patient, with the Aff(ected) primitive; the By primitive for Means; and the For primitive for Reason), and divides into a Thematic (spatial) tier and an Action (causal) tier, with the stipulation that an entity can have at most one role from each tier. However this proposal is never fully elaborated.

To return to the grammatical predictions that Jackendoff makes on the basis of roles, he states two Thematic Hierarchy Conditions, governing the formation of the passive, and of reflexives (Jackendoff, 1972, pp.43,148), based on a single Thematic Hierarchy (2.115).

(2.115) The Thematic Hierarchy

1. Agent
2. Location, Source, Goal
3. Theme

The condition on passives is given below (2.116).

(2.116) Thematic Hierarchy Condition on Passives
The passive *by*-phrase must be higher on the Thematic Hierarchy than the derived subject.

Measure objects of verbs such as *cost* and *weigh* bear the role Location (evidenced by the prepositions *in* and *at* in 2.117 and 2.118), and so their subjects are Themes (due to the stipulation that a Theme is obligatory in every clause). A passive using these verbs (2.119, 2.120) results in a Theme *by*-phrase and a Location subject, violating the hierarchy condition. Jackendoff claims that ambiguous active sentences, where the subject is Theme, and optionally Agent (e.g if *John* is not inadvertently touching the bookcase in (2.121), are only acceptable in the passive (2.122) with an agentive reading.[17]

(2.117) The champ weighed in at 654 pounds

(2.118) I wouldn't buy oranges at 25 cents apiece

(2.119) *Five dollars are cost by the book

(2.120) *Two hundred pounds are weighed by Bill

(2.121) John was touching the bookcase

[17]The author does not agree with this judgement. The passive seems to be marginally acceptability, and neutral between the agentive and non-agentive readings.

(2.122) The bookcase was being touched by John

Jackendoff does not address which of its multiple roles determine an entity's place in the hierarchy, nor whether the rule applies to passives that lack a *by*-phrase, but the condition could easily be modified to account for these cases.

A more serious problem is posed by inanimate agents, and animate non-agents. The passive (2.123) is perfectly acceptable, but *camper van* (which is inanimate, and so cannot be an Agent) appears to be a Theme, violating condition (2.116). Presumably Jackendoff would consider both *the public* and *Dougal* in (2.124, 2.125) to be Location, because of the existence of paraphrases using *lovable to*. Accordingly, condition (2.116) would predict that neither would passivise successfully, but this is not the case. An animate passive subject (2.126) appears more acceptable than an inanimate passive subject (2.127), despite both sharing the same role in Jackendoff's system.

(2.123) Our geraniums were crushed by the camper van

(2.124) The public loves David Hasselhoff

(2.125) Dougal loves cream-cake

(2.126) David Hasselhoff is loved by the public

(2.127) *Cream-cake is loved by Dougal

On reflexives, Jackendoff first asserts a syntactic restriction that the antecedent must appear in an external (i.e. subject) position, and the reflexive pronoun in an internal (i.e. post-verb) position, hence the unacceptability in most English dialects of (2.129).[18]

(2.128) John$_i$ shaved himself$_i$

(2.129) *Himself$_i$ shaved John$_i$

To the syntactic prescription, he adds the semantic constraint in (2.130).

(2.130) Thematic Hierarchy Condition on Reflexives
A reflexive may not be higher on the Thematic Hierarchy than its antecedent.

Since the antecedent *John* is an Agent, and the reflexive *himself* a Theme, (2.128) does not violate the hierarchy condition and is valid. However, if the sentence is passivised, either the syntactic constraint must be violated by putting the reflexive in subject position (2.131), or the hierarchy condition violated by having an Agent reflexive (2.132). Similarly to (2.122), Jackendoff claims that a verb such as *touch*, which can have a volitional or inadvertent reading, is only acceptable in an agentive reading with a reflexive pronoun (2.133) (where the condition on reflexives is not violated).

(2.131) *Himself$_i$ was shaved by John$_i$

(2.132) *John$_i$ was shaved by himself$_i$

(2.133) John$_i$ was touching himself$_i$

[18]Objects can also be antecedents to more oblique reflexives, e.g. *The psychiatrist introduced Mary to herself*. And reflexive pronouns can be used as emphatic variants of personal pronouns in any argument position in Irish English – (2.129) would be interpreted as "*He*$_{\text{i}}$ shaved John$_{\text{j}}$".

Of course the condition on reflexives is undermined somewhat, since a reflexive and its antecedent share the same referent, and it is not obvious how its separate syntactic identities can be disentangled semantically. Ignoring syntactic differences, the deliberately caused effect to oneself in shaving (with two role relationships in Jackendoff's analysis) is much like that of rolling oneself down a hill (2.108) which is analysed as having a single role.

Turning to psychological predicates, Jackendoff analyses the sentences below as having *Bill* as Theme (since he is attributed with the characteristic of pomposity), and *pompous* as an abstract Location (in a similar manner to *angry* in 2.102). Since something may be *striking to* somebody else, *Harry* is a Goal in (2.134), and by "semantic parallelism", also a Goal in (2.135) (Jackendoff, 1972, p.45). Thus, the condition on the reflexive would rule out (2.136) and allow (2.137). However, the marginal acceptability of (2.136) may be due to implausibility – it is hard to see how someone could hold an initial or superficial opinion about their own personality, as *strike* implies. The more plausible *Bill struck himself as clever* is more acceptable.

(2.134) Bill strikes Harry as pompous

(2.135) Harry regards Bill as pompous

(2.136) ?Bill strikes himself as pompous

(2.137) Harry regards himself as pompous

In general, the hierarchy conditions can present problems, due to the differing fashion in which Jackendoff defines his roles. The spatial roles of Location, Source and Goal are structurally marked by prepositions such as *at, from* and *to*, either in sentences themselves, or in closely related paraphrases. Agent, on the other hand, is defined semantically, in that it is volitionally involved, and as a result, must be animate. Theme has the structural characteristic of having to be present, and semantically of undergoing a transition or having a state attributed to it.

Considering the verbs *please* and *like*, that take an Experiencer and Stimulus arguments, but in different positions, it is not easy to decide on a role analysis. It is possible to find *pleasure in* something, suggesting that the Stimulus is a Location, and the Experiencer is a Theme. On this analysis, a reflexive pronoun should not occur with *please*, since as a Location, it would be higher than its Theme antecedent, but as (2.138) and (2.139) illustrate, such uses are attested. Alternatively, since something can be *pleasing to* someone (this is Jackendoff's analysis), the Experiencer can be considered a Goal, with the Theme role going to the Stimulus. This predicts that (2.138) and (2.139) are acceptable. However, by semantic parallelism (Experiencer=Goal, Stimulus=Theme), one would expect sentences such as (2.140) to be ruled out.

(2.138) Oh well, please yourself, ducky.[19]

(2.139) Dorothy, needed a break. She was going home to her flat to please herself for a few weeks.[20]

(2.140) I startle Dudley, and myself, by choosing this moment to ask him if he likes himself.[21]

[19] BNC text AC5, Fiction, "Paper Faces" by Rachel Anderson.
[20] BNC text CBS, Fiction, "The Fifth Child" by Doris Lessing.
[21] BNC text ECT, Feature Journalism, *Esquire* magazine.

Taking a sentence such as (2.141) *himself* can be analysed as a Location, and *he* as a Theme. That *deliberately* can be inserted acceptably indicates that the sentence has an agentive reading. Jackendoff does not explain how to decide which of the two animate arguments present is the Agent, but the acceptability of the sentence, together with his condition on the reflexive indicates that it must be *he*. By this analysis though, a sentence such as (2.142) should be impossible, since *she* is explicitly not Agentive.

(2.141) He ... turned back on himself, sprinted in big circles and then wee circles, all with the ecstatic foal in hot pursuit.[22]

(2.142) She got lost and turned back on herself without realising.

A final acceptable and attested example that unambiguously violates Jackendoff's condition on the reflexive is (2.143) – *itself* must be considered a Location, due to the preposition *on*, and *trail* must then fill the role of Theme. Since it is not animate, it cannot be additionally Agent. Hence a Location reflexive pronoun outranks its Theme antecedent, contrary to (2.130). And while it would seem a reasonable suggestion to extend the notion of Agent to include inanimate but active participants, such as *storm* in (2.144), including cases such as *trail* within the ambit of Agent would seem to stretch the notion to vacuity.

(2.143) The trail turns back on itself at a switchback, and you'll recross the hill at a lower elevation, and then reemerge into sunny chaparral.[23]

(2.144) The storm blew itself out

In sum, the essential problem with localist approaches to argument structure description is that a single meaning can be expressed in forms that differ in the localist structure they suggest, resulting in similar problems of decidability as were found in atomic approaches. As just noted, in English *someone can find pleasure in something* suggesting that the Stimulus is a Location and the Experiencer the Theme, while the fact that *something can be pleasing to someone* suggests the opposite analysis. In Irish, a cognitive state can be a Theme (2.145) or a Location (2.146).

(2.145) Tá fearg orm
is anger on-me
'I am angry'

(2.146) Tá mé ar buile
is me on frenzy
'I am furious'

2.5 Decompositional Approaches

Predicate decompositional approaches often posit a primitive predicate for causation (Levin and Rappaport Hovav, 2005, p.69), as was seen in an early example of Jackendoff's LCS (2.86), rather than dealing in terms of Agents, Actors and Instruments. Croft (1993) investigates verbs that describe mental states or cognitive activity ('psych verbs'). In his view, verbs that take a Stimulus subject

[22] Children's Fiction, "Maxi's Angels": http://www.maxisangels.co.uk/chapter_10.php, viewed 21st July 2006.

[23] Hiking Guide, "El Sereno Open Space Preserve": http://www.bahiker.com/southbayhikes/elsereno.html, viewed 21st July 2006.

(such as *amuse, disgust, frighten*) include a Cause predicate in their lexical semantics, and equivalent verbs that take an Experiencer subject (such as *enjoy, loath, fear*) do not. While a causative verb such as *please* can take a means phrase (*by coming in ...* in 2.147), the stative *like* cannot (2.148, 2.149) (*ibid*, p.57):

(2.147) John pleased his boss by coming in early every day

(2.148) *John's boss likes him by coming in early every day

(2.149) *John was liked by his boss by coming in early every day

While this account certainly captures differences between the two classes of Experiencer subject and Stimulus subject verbs, the question is complicated by apparently stative uses of "causative" verbs (2.150) (cf. Levin and Rappaport Hovav, 2005, p.56), and causative uses of "stative" verbs (2.151).

(2.150) a. Multinationals have in the past been checked by laws which require Indians to take majority shareholdings in foreign companies operating in India. These regulations have, for instance, forced IBM out of India altogether. But the latest scheme to boost exports and control imports does not **please** India's software industry.[24]

b. ... India's software industry does not **like** the latest scheme to boost exports and control imports ...

(2.151) a. I've always **liked** the rain, the city at night, wet streets stretching into winter darkness, a peculiar feeling of freedom that it contains.[25]

b. The rain, the city at night, wet streets stretching into winter darkness, a peculiar feeling of freedom that it contains have always **pleased** me.

The *swarm* (Partee, 1965; Anderson, 1971) and *spray/load* diatheses (Rappaport and Levin, 1988) are often viewed as encoding differences in meaning, in the degree of affectedness that they attribute to their arguments. Anderson's analysis of *swarm* sentences (2.47 on page 33), which viewed the direct object as encoding "holistic" affectedness, and its oblique position as encoding "partitive" (partial) affectedness, could be equally applied to sentences such as (2.152) (Rappaport and Levin, 1988, p.26). Rappaport and Levin themselves prefer an analysis based on causation in which (2.152a) describes a caused change of location (the *hay* moves) and (2.152b) describes a change of state (the *wagon* becomes full). These two forms can be viewed as separate senses of the verb *load*, the first being roughly synonymous with *put* and the second closer to *fill*.

(2.152) a. John loaded hay onto the wagon.

b. John loaded the wagon with hay.

These accounts aim to explain anomalies in the sentences (2.153c) or (2.154c) – the total affectedness of the *hay* in the first should be incompatible with a bale being left over, and the affectedness of the wagon in the second should be inconsistent with room being left over (Beavers, 2004).

(2.153) a. John loaded the hay onto the wagon, filling the entire wagon.

b. John loaded the hay onto the wagon, and still had room left over.

c. #John loaded the hay onto the wagon, and still had a bale left over.

[24] BNC text B7H: *New Scientist* magazine, 1991.
[25] BNC text HTW: Fiction, *The eagle has flown*, Jack Higgins, 1991.

(2.154) a. John loaded the wagon with the hay, leaving none behind.

b. John loaded the wagon with the hay, and still had a bale left over.

c. #John loaded the wagon with the hay, and still had room left over

While the illformedness of these examples may be open to question[26], any suggestion of increased affectedness can be outweighed by pragmatic factors. In the constructed example (2.155a), *the walls* can be realised as a direct object to emphasise the mess made, without entailing that the wall is pasted in its entirety (see Dowty, 1991, pp.587-592 for a similar point on the effect of real-world knowledge on the interpretation of this structure).

(2.155) a. The toddler sprayed the walls with his semolina, but the patch behind him remained unscathed

b. The toddler sprayed his semolina all over the walls, but still has some left over to throw at his mother

Definite singular generic (*the pesticide* in 2.156) and unquantified plural references (*pentagrams* in 2.157) can appear in the direct object position, even though in both cases there is no bounded referent which could be seen as holistically affected. In both (2.158) and (2.159) instructions are given on how to spray a quantity of liquid from a container, but in neither case does it seem plausible that the intention is to use the entire contents, as these theories would predict.

(2.156) EPA's decision would make the acceptable exposure level of the pesticide one-tenth of what's currently allowed, the Post said. Farmers will still be allowed **to spray the pesticide on crops**, but its agricultural use will be reduced to a degree not yet decided.[27]

(2.157) ... we drove up to Tacoma and scored some crack and went to a graveyard where we smoked crack and took mescalin and **spray painted pentagrams on mausoleums** ...[28]

(2.158) just spray/soak the area really really good with hairspray (the cheap kinds work best), then rinse with cold water, then spray it again, repeat until it's gone or almost gone. the ink will start to spread out when **you spray the hairspray on it**, but don't worry it will eventually rinse out.[29]

(2.159) The water from the spray bottle beads on top of the cream and acts as the lubricant. [...] Do not **spray the water on anything but your slide** ...[30]

A context can be constructed in which the discourse constraint that tries to place given elements earlier in the sentence than new elements can overrule the semantic constraint that tries to place wholly affected participants in the direct object position. In (2.160a) *hay* is already the topic of conversation, a given element, and so fits more comfortably before the new constituent *wagon*, than after it in (2.160b).

[26] Both are perfectly wellformed in my opinion.

[27] Associated Press story on environmental warning; reposted to a discussion forum on keeping skunks as pets, viewed August 2006 [http://www.rdwarf.com/pipermail/skunks/2000-June/003016.html]. Repeated from page 5.

[28] Posting to discussion forum, viewed August 2006, [http://www.satanosphere.com/story/2001/6/29/145857/275].

[29] Posting to discussion forum, concerning removing ink from clothes, viewed August 2006, [http://www.subkultures.net/tonalwar?readjid=1896574].

[30] Instructions on trombone maintenance from school band website, viewed August 2006, [http://www.centennial.k12.mn.us/cms/dept/music/band/presources/tbnslide.htm].

(2.160) a. A: Where is all the hay gone?
B: John loaded it onto the wagon, but still had a bale left over, so he put that in the shed.

b. A: Where is all the hay gone?
?B: John loaded the wagon with it, but still had a bale left over, so he put that in the shed.

Pinker (1989) gives a similar account of differences between the two forms of the dative diathesis. In his analysis the double object construction entails the Have predicate (2.161a) while the prepositional object variant (2.161b) entails only movement (though possession is not ruled out).

(2.161) a. Ann gave Beth the car
[$_{\text{EVENT}}$ give [Ann Beth [$_{\text{STATE}}$ HAVE Beth the car]]]

b. Ann gave the car to Beth
[$_{\text{EVENT}}$give [Ann the car [$_{\text{EVENT}}$ GO the car [$_{\text{PATH}}$ to [$_{\text{PLACE}}$ Beth]]]]]

The presence of the Have predicate in this form explains for Pinker (1989, p.110-111) and others (e.g. Krifka, 2004) why verbs of movement only appear in the prepositional object form (2.163), why locations which are not recipients cannot take the double object configuration (2.162), and why metaphorical uses (which do not involvement movement) appear only in the double object configuration (2.164).

(2.162) a. Tom sent Bob a package

b. *Tom sent Dublin a package

(2.163) a. I lowered the box to John

b. *I lowered John the box

(2.164) a. *The lighting here gives a headache to me

b. The lighting here gives me a headache

But attested examples can be found that use the double-object variant when receipt (2.165) or possession (2.167) is not achieved. Bresnan and Nikitina (2003, pp.6, 8) also provide counter-examples to the claim that motion verbs cannot take the double object (2.168) and that metaphorical uses cannot take the prepositional object (2.169).

(2.165) Grant Mann would have been 31 last Friday. **His family sent him birthday cards**, but he never received them, he was already dead.[31]

(2.166) You were sent instructions to evacuate but your own computer equipment was unable to receive our instructions and advice, as you very well know.[32]

(2.167) I will promise you this: if you **send me something to give away**, I'll be sure to send you a photograph of the child you helped[33]

(2.168) ... Buddha lowered him the silver thread of a spider ...

[31] BNC text K1K: Central Television news script.
[32] BNC text BN7: Leaflet on health issues
[33] *AfricanKelli* Weblog, 7th June 2005, http://africankelli.blog.com/2005/6/.

(2.169) ... guaranteed to give a headache to anyone who looks hard at the small print.

These predicate analyses of the dative also depend on the verb having more than one sense. Bresnan and Nikitina (2003, p.5) challenge this view with examples such as (2.170). It seems implausible that the character J.C. is choosing not to specify possession of the doughnut in the first use of *carry*, but does in the second.

(2.170) "You **carrying a doughnut to your aunt** again this morning?" J.C. sneered. Shelton nodded and turned his attention to a tiny TV where "Hawaii Five-O" flickered out into the darkness of the little booth. "Looks like you **carry her some breakfast** every morning."

As Levin and Rappaport Hovav (1996) point out, the challenge is "how to provide an analysis that balances the similarities and differences in meaning" in diatheses, but these rigid models based on predicate forms do not appear to be able to account for the variation seen in real language usage. One reason for this may be the emphasis on the lexeme internal semantic of verbs, at the expense of verb-argument relations.

2.6 Prototype Approaches

Dowty (1991) begins his account of semantic roles and linking by considering the limitations of previous approaches. He points out that for classical inventories of roles (by which he means Fillmore, 1968 and theories that built on it) to function properly, the semantics associated with verb argument positions must have certain properties:

> "... the meanings of all natural-language predicates must turn out to be of a very particular sort: for every verb in the language, what the verb semantically entails about each of its arguments must permit us to assign the argument, clearly and definitely, to some official semantic role or other – it cannot be permitted to hover over two roles, or to 'fall in the cracks' between roles" Dowty (1991, p.549)

He recognises the notion of semantic roles as a cognitive phenomenon, but thinks it should be justified and explained "in terms of its referential semantic significance." For him verb specific roles, such as those used in HPSG, ("individual thematic role" in his terminology) are defined by the set of all properties that hold of a particular argument position of a particular verb. For example the Builder role is "the set of all the things you can conclude about *x* solely from knowing that the sentence *x builds y* is true". A generalised role (a "thematic role type") is the intersection of all the properties of a set of verb-specific roles. However, as he points out, linguists are not interested in any arbitrary collection of individual verb roles, but rather a particular set that "plays a special role in linguistic theory" Dowty (1989, pp.73-82), what he calls "L-Thematic Role Types". Given that this is what is being investigated, he suggests three requirements to be adhered to when assembling a set of roles:

(2.171) (Completeness) every individual thematic role contains some L-thematic role type (or as we may equivalently say, every argument position of every verb is "assigned" an L-thematic role type).

(2.172) (Distinctness) Every argument-position of every verb is distinguished from every other argument-position of the same verb by the L-thematic role types the two argument-positions are assigned.

(2.173) (Independence) The properties in an L-thematic role type must be characterizable independently of the relations (denoted by natural language verbs) that entail them.

He thinks it unlikely that any theory of roles can perform adequately both argument linking functions (as suggested by Fillmore and others), and argument indexing functions in formal models, like Chomskyan theories (Theta Theory, see p.36), Role and Reference Grammar (RRG, van Valin and LaPolla, 1997) and Lexical Functional Grammar (Kaplan and Bresnan, 1982). Though some theories had moved towards individual roles, they did not allow syntactic generalisations. What was needed was a small number of roles with both coverage and distinctness.

Dowty (1991, p.563) specifically excludes notions of perspective from his theory of linking: "I want to suggest that we rule out such perspective-dependent notions as Figure/Ground ... as candidates for semantic roles. This is not to deny the existence of these distinctions or their importance, but to propose only that thematic role is the wrong rubric for them."

For the purposes of argument selection, he proposes that only two proto-roles are needed – Agent and Patient – and that these are "cluster concepts" rather than discrete (i.e. atomic) types (Dowty, 1989, p.111), very similar to RRG's 'macroroles' of Actor and Undergoer.[34] These are non-primitive concepts that are characterised by the verbal entailments towards argument positions with which they are compatible (1991, p.547): "an argument of a verb may bear either of the two proto-roles (or both) to varying degrees, according to the number of entailments of each kind the verb gives it". His entailments, in large part inspired by the qualities of the universal subject of Keenan (1976), are as follows:[35]

(2.174) Contributing properties for the Agent Proto-Role:

a. volitional involvement in the event or state
b. sentience (and/or perception)
c. causing an event or change of state in another participant
d. movement (relative to the position of another participant)
e. (exists independently of the event named by the verb)

(2.175) Contributing properties for the Patient Proto-Role:

a. undergoes change of state
b. incremental theme
c. causally affected by another participant
d. stationary relative to movement of another participant
e. (does not exist independently of the event, or not at all)

His ProtoAgent properties include a broad definition of volition, including decisions not to act (such as *ignore*, or *prevent*). By sentience he means sentience relative to the event in question (i.e. awareness) and recognises that its boundary is clouded by cases such as computers and intelligent animals.

[34] RRG distinguishes between the many thematic relations, such as Speaker, Judgement, Possessed, that "have no independent status, [being] really just mnemonics for the argument positions", and the two 'macroroles' of Actor and Undergoer that determine the realisation of argument structure via a hierarchy in terms of predicate configurations in RRG's formal representation, 'logical structure' (van Valin, unpublished, pp.14-16).

[35] Dowty's entailments share much with Fillmore's (1977b) saliency hierarchy – see (2.183) on page 58.

His notions of causation and movement may be graded also – for example moving one's own arm is not as strongly causal for him as moving something external to oneself. The final ProtoAgent property of independent existence (i.e. not being created or destroyed as a result of the event) is questionable as he is not sure to what extent this is a feature of perspective or of semantics (*ibid.* pp.572-573).

Several of the ProtoPatient properties are the converse of ProtoAgent properties. Dowty explains that this duplication is necessary to account for selection of the direct object in verbs with more than two arguments. His feature of incremental theme is based on the insight of Verkuyl (1993) that direct objects can have a "measuring out" function in determining the aspectual nature of a verb, and this comes into play later to account for the distribution of the *load/spray* and related diatheses (*ibid.* pp.567,587). Many traditional roles can be expressed in terms of his entailments. For example, an Instrument entails causation and movement, without sentience or volition. Many of the aspects of role hierarchies follow naturally from these correspondences, though his features do not incorporate any aspects of the localist Source and Goal.

His argument selection principles are as follows (*ibid.* p.576):

(2.176) a. Argument Selection Principle: In predicates with grammatical subject and object, the argument for which the predicate entails the greatest number of ProtoAgent properties will be lexicalized as the subject of the predicate; the argument having the greatest number of ProtoPatient entailments will be lexicalized as the direct object.

b. Corollary 1: If two arguments of a relation have (approximately) equal numbers of entailed ProtoAgent and ProtoPatient properties, then either or both may be lexicalized as the subject (and similarly for objects).

c. Corollary 2: With a three-place predicate, the nonsubject argument having the greater number of entailed ProtoPatient properties will be lexicalized as the direct object and the nonsubject argument having fewer entailed ProtoPatient properties will be lexicalized as an oblique or prepositional object (and if two nonsubject arguments have approximately equal numbers of entailed P-Patient properties, either or both may be lexicalized as direct object).

d. Nondiscreteness: Proto-roles, obviously, do not classify arguments exhaustively (some arguments have neither role) or uniquely (some arguments may share the same role) or discretely (some arguments could qualify partially but equally for both proto-roles).

Dowty concedes that this model cannot account for the small number of verbs such as *undergo, suffer, inherit,* and *receive* that place ProtoPatients in subject position (*ibid.* p.581-583).

He explains the existence of symmetrical predicate pairs such as *fear/scare*, *like/please* and *strike/regard* by the fact that their arguments have balanced numbers or proto-entailments, and so are not strongly inclined to be realised in either configuration. Verbs that are optionally asymmetrical like *kiss* (the recipient may or may not be willing) can be accounted for. If one of the participants (*Diana* here) is the instigator, s/he is more agent-like and a structure such as *Diana kissed/met Doug* results. If both participants are equally active, Corollary 1 allows for both *Diana kissed/met Doug* and *Diana and Doug kissed/met.*

But the same principle might falsely predict that *Diana scares Doug* could be paraphrased to *Diana and Doug scare*. The two arguments of verbs such as *resemble* are identical in terms of proto-entailments, so *A resembles B* should be fully interchangeable with *B resembles A* – but consider *That politician looks like a crook*, and **A crook looks like that politician.*

The dative and benefactive diatheses also present problems. The wording of Corollary 1 is intentionally vague. If the Recipient and the Theme are considered to have "(approximately) equal numbers" of ProtoPatient entailments, then the paraphrases *give something to Mary* and *give Mary something* are both correctly predicted, but so is *@give something and Mary*.[36] If on the other hand Corollary 1 is interpreted more restrictively, and consider the Theme as the better ProtoPatient, then only *give something to Mary* would be predicted, and *give Mary something* would be ruled out.

One other possible criticism of Dowty is his choice of terms. Since he largely restricts himself to English in his discussion, it might be suggested that what he is describing are proto-subject and proto-direct object. His explanation is that for his theory to be able to generalise to non-nominative languages, the independent categories of agent and patient must be retained. Ergative languages, in contrast to nominative languages, choose the proto-patient as their "syntactic pivot" (*ibid.* p.582) (what a syntactic universalist might refer to as a subject).

Dowty's initial presentation of entailments is very strong – these are analytic implications (also termed "lexical entailments") of a particular predicate towards each of its arguments, that "follow from the meaning of the predicate in question alone". If this is applied strictly, it rules out what may seem like reasonable features for many argument positions. For example, it means that volition cannot hold of the subject of the verb *murder*, because of the possibility of a person being coerced into murdering another. While he does entertain the possibility that his entailments might be weighted, or their satisfaction might be graded, this is not developed (Dowty, 1991).

Finally, Dowty's attempt to exclude perspective from his account of argument structure, while initially viewed as simplifying matters, leads to problems. He is convinced that "Thematic roles should all have event-dependent definitions, never discourse-dependent ... ones" (1991, p.613), and goes on to conclude that discourse notions should not play a part in argument selection. Despite this, the movement entailments for ProtoPatient and ProtoAgent presuppose the existence of the figure/ground distinction of Talmy (1985) that Dowty disavows: if a single event participant is moving, then all other event participants are moving relative to it. Thus the entailment applies to all the arguments and ceases to be a distinguishing feature. To have a distinction between static and moving participants in a single event requires an external background reference.

Dowty chose not to incorporate the saliency hierarchy of Fillmore (1977b) in his theory, but he does concede that subject is a weak indicator of topic (1991, p.564). This contributes to his failure to adequately account for the choice between *buy/sell*, in which the vendor and buyer have similar claims on Subjecthood. It also fails to predict the choice of dative diathesis form (where recipients and patients are similarly good candidates for the direct object), or the passive construction (in which the ProtoAgent is linked to an oblique position, and the ProtoPatient appears as subject).

2.6.1 Derivative Approaches

Baker's (1997) approach to linking is decompositional in nature, and introduces the ideas of proto-roles to Chomskyan Principles and Parameters theory (Chomsky, 1981). He appeals to three proto-

[36] The at symbol '@' is used to mark sentences that are semantically divergent from their counterparts (i.e. for structurally related sentence pairs that are *not* paraphrases). See the appendix (page 183) for a list of marks and other notations used.

roles in his theory of "agent/causer, theme/patient and goal/path/location" (Baker, 1997, p.38), and hypothesises that linking patterns are universal at deep structure:

(2.177) The Uniformity of Theta Assignment Hypothesis (UTAH)
Identical thematic relationships between items are represented by identical structural relationships between those items at the level of D-structure

The nature of Chomsky's theories, in which covert structures have formal reality, and movement operations are widely applied, allows Baker to maintain this hypothesis. For example, the sentences in (2.178) are synonymous, and so are assumed to have identical deep structure. Since their surface structure is different, one or other of them must be generated by some movement operation.

(2.178) a. Mary worries about the ozone layer

b. The ozone layer worries Mary

Similarly an ergative language can have the same deep structure as a nominative one. In (2.179) the deep structure of agent-verb-patient which is realised directly in the surface structure of a nominative language, illustrated by (2.179a), is maintained in an ergative language through the device of a co-indexed trace (2.179b).

(2.179) a. [The man] hit [the woman]

b. [The woman]$_i$ [the man] hit t_i

Reinhart also aims to "implement as many of [Dowty's] intuitions as [she] can" (2002, p.9) in her Chomskyan account. Chomskyan systems require features to be formally coded, so that they are generally accessible to the system. She proposes two features (*ibid.* pp.5-13): /c is "cause change" and a "perceived sufficient condition"; /m is a "relevant mental state". Clusters of these two features characterise the semantic role of verbal arguments, and many of these clusters correspond to more traditional roles:

(2.180) a. [+c+m] agent

b. [+c-m] instrument

c. [-c+m] experiencer

d. [-c-m] theme/patient

e. [+c] cause

f. [+m] sentient

g. [-m] subject matter/locative source (Typically Oblique)

h. [-c] goal/benefactor (Typically Dative (or PP))

i. [ ] - Arb(itrary)

Cluster features can express multiple frame configurations economically. For example the subject of *destroy* specifies the [+c] cluster, which neatly encompasses Agents, Actors and Instruments. For her, these features are formal, but conceptually based on real-world knowledge. She claims that their formal simplicity resolves the issue of argument structure learnability (Pinker, 1989), but claims that they nevertheless have a certain independence from external reality.

Reinhart mentions the unergative verb *glow* (e.g. *the diamond glowed*) as an example of a [+c] subject verb, that is not internally causing – it needs an external source of light to produce the glow, and so by semantics alone, would be expected to be unaccusative (Reinhart, 2002, p.7). This makes one wonder how these features can be expected to aid in learning, if they can be arbitrary and independent of reality. A more compelling answer is perhaps that in people's perception, diamonds appear to glow from within and so are [+c]. This is the approach Haiden (2005, p.273) takes when he further grounds Reinhart's work in the ideas of *folk psychology* or *folk physics*, and sees the /c feature as representing *force* and the /m feature *sentience*.

Sanfilippo (1990) again builds on Dowty, this time introducing two additional syntacto-semantic roles: prepositional and propositional. His framework is of neo-Davidsonian semantics inside a phrase structure grammar formalism (Unificational Categorial Grammar). The system lacks grammatical function, using ordering relations in its subcategorisation features instead. Hence the linking procedures work in terms of symbol configurations: the P-Agent is tied to the "innermost sign" and the P-Patient to "next innermost sign". For rigid argument structure languages such as English, the Role Association rule is essentially identical to Chomsky's Theta Criterion:

(2.181) Role Association *English*
"All syntactic arguments of a predicate must be linked to a unique role, and all entailed roles to a unique syntactic argument"

While for an free argument dropping language such as Japanese, no specification of role association is required at all (Sanfilippo, 1990, pp.335-339).

Davis (1996) also adopts ideas of Dowty's into a phrase structure grammar, and again ordering in the syntactic structure plays the role of grammatical function. He concentrates on various syntacto-semantic classes of verbs that share valency frames[37] and on how a wide range of patterns can be expressed economically with a hierarchy of sorts in the lexicon. For example his ditransitive verb frame inherits most of its features from the monotransitive frame. Each linking sort "partly defines the mapping between the values of proto-role attributes in the semantics and the ordered elements of the ARG-S list that indicate a predicator's syntactic requirements" (Davis, 1996, p.183). His proto-role attributes are ACT (Agent), UND (Patient), FIG and GRD (Talmy's Figure and Ground; 1985). To these he adds an EFFECT feature that describes resultant states. For example in dative construction, he includes the GRD element (the Recipient) in EFFECT to signal that possession is as a result of the event (*ibid.* pp.121,147) (he generally treats Possessors as Ground and Possessions as Figure).

Finally, Primus (2002; 2006) adopts some of Dowty's ideas to provide an Optimality Theory account of argument linking. She defines the ProtoAgent as an argument *x* that typically stands in the following predications relative to the situation described *s*, and another participant in the situation *y* (2006, p.3):

ctrl(x,s) *x* causes (controls) the situation *s*

exp(x,y) *x* is aware of (experiences) the participant *y*

phy(y) *y* is affected

poss(x,y) *x* possesses participant *y*

[37] Thus his approach is compatible with listing approaches such as Levin (1993) and Dixon (1991); the view of Goldberg (1995; 2004) that frames aid learning; and the use of frames as criteria for making verb sense distinctions Palmer (2000).

ProtoPatient is defined similarly as typically being the participant *y* in these predications. In the case of situations with more than two participants, an argument that is both less prototypically agent-like than the most agent-like participant and is less patient-like than the most patient-like argument is termed the Proto-Recipient. In this way Primus sees the ProtoPatient as being causally dependent on the ProtoAgent, and the Proto-Recipient as being causally dependent on both ProtoAgent and Patient. However, it could be equally argued that an Agent is dependent on a Patient, since an Agent needs something to act on. Utterances such as *Charles fell off the cliff* that appear to have a Patient but no Agent also challenge her view.

In sum, with the possible exception of Davis (1996) the main contribution of these derivative approaches has been to make Dowty's ideas compatible with a pre-existing formalism. They do not radically increase the range of argument structuring phenomena that can be explained, or provide a deeper grounding of the semantic notions at play.

2.7 Discourse Approaches

The preceding sections have described the considerable literature that has considered what and how aspects of the situational meaning may be expressed in differences in argument structure. It is perhaps less obvious that elements of pragmatics, of discourse structure and of text structure, are determining factors in argument linking. While verb-argument semantics are the primary driver of argument selection, information structure (particularly the given/new and topic/comment distinctions[38] – see e.g. Prince, 1981; Halliday, 1985, pp.30-33; Vallduvi and Engdahl, 1996), topical saliency and argument constituent weight (i.e. length or structural complexity of the argument phrase) play a role in argument ordering. It has been long recognised that the grammatical subject is associated with topicality and definiteness (see e.g. Keenan, 1976; Croft, 1991, p.151), and the figure/ground distinction has been invoked to explain diathesis choices (Croft, 1993). In English, given constituents are likely to be realised earlier in the clause than new constituents, and light constituents earlier than heavy ones (Arnold et al., 2000). This may be achieved by use of alternative verb choice (e.g. *fear/frighten*, *contain/comprise*, *give/get*) or by the use of the passive (Biber et al., 1999, pp.940-941). For some other languages, such as Tagalog, the subject is chosen primarily on the basis of focus relations (Bisang, 2006).[39]

Kuno (1976, p.431-434) proposed that the degree of empathy or identification the speaker feels with event participants plays a role in argument structure realisation, and he pointed out relationships between empathy (identification might have been a better term, but that might cause confusion with concepts of semantic reference) and both grammatical role, and givenness. His empathy hierarchy is very straightforward (2.182) – a speaker is most prone to identify with themselves, followed by the listener, followed by third parties – and is in effect an anthropocentric saliency hierarchy.

(2.182) Speaker > Hearer > Third Person

In some languages this constraint is obligatory. Bresnan et al. (2001) describe the North American language Lummi, in which subjects cannot be lower on this person hierarchy than objects – a sentence such as *She likes me* would not be licensed, and would have to be realised as a passive.

[38] Authors have used a wide range of terminologies to describe differences in discourse prominence, including theme/rheme, topic/comment, background/Kontrast, figure/ground, focused/unfocused, without reaching a consensus (Kruijff-Korbayová and Steedman, 2003). In this work the more tangible distinction of given/new (see e.g. Prince, 1981) is appealed to.

[39] Dryer (1997) provides a slightly different analysis of the closely related language Cebuano. In his view there are two subject-like grammatical relations operating in parallel – one that signals focus and the other that marks an actor.

Fillmore considered the effect of saliency and perspective on argument structure in (Fillmore, 1977a,b). While a speaker presumably conceptualises complete scenes, they may select a particular perspective on an event, choosing to express certain aspects and omit others. In turn, verbs may express a particular perspective – e.g. *buy* necessarily brings the customer into focus as subject, and *sell* the salesperson. Fillmore (1977a) introduced the notion of saliency to explain choice of perspective. According to his theory, it is the most salient participants in an event that are realised as *nuclear elements* (i.e. subject and direct object – he states explicitly that his analysis cannot account for the distribution of the the indirect object). Without placing them in a hierarchy, he lists the relevant properties of saliency as humanness and animacy; motion and change of state; definiteness; and totality of affectedness. A case hierarchy (again not formalised, but presumably along the lines of (2.32) – see section 2.3) then selects the subject in the case of more than one noun phrase being nuclear. Fillmore (1977b, pp.91-109) goes further in suggesting that the case hierarchy may be discarded entirely, in favour of a saliency hierarchy which combines elements of semantics and pragmatics (though, in his words, it remains "sheer speculation"). This hierarchy (2.183) selects the highest ranked element as subject of a sentence, and then, if the verb's subcategorisation features allow it, the next-highest-ranked element as direct object.

(2.183)
- An active element outranks an inactive element.
- A causal element outranks a noncausal element.
- A human (or animate) experiencer outranks other elements.
- A changed element outranks a nonchanged element.
- A complete or individuated element outranks a part of an element.
- A 'figure' outranks a'ground.'
- A 'definite' element outranks an 'indefinite' element.

Dell et al. (1993, p.530) also describe concepts that are more animate, substantive or imageable as having privileged access to prominent grammatical positions. Croft (1991, p.151) outlined the "natural topic hierarchy" (2.184) below. The relative prominence of clause arguments on this hierarchy would map isomorphically onto the hierarchy of grammatical relations (2.185).

(2.184) 1st/2nd person > 3rd person > Proper Noun > Human > Animate > Inanimate

(2.185) Subject > Object > Oblique

More recently, Aissen (2003) has suggested a model in which animacy and definiteness determine object selection, and Wunderlich (2006) has suggested a model of argument selection that takes account of "sortal (or referential) salience ... such as person, animacy or specificity" and "informational salience ... such as topic and focus". Bresnan and Nikitina (2003, p.34) also appeal to a "hierarchy of person or speech act participants" to explain the dative diathesis.

In sum, while it is unlikely that linking patterns can be explained exclusively by appeal to discourse notions, it may be difficult to provide an adequate description without taking these into account.

2.8 Conclusions

Atomic approaches like that of Fillmore (1968) pose obvious, and apparently insurmountable problems of decidability – how to decide what roles to include in a schema, and how to decide among

roles when annotating individual arguments. While localist approaches capture many interesting generalisations about the manner of expression of argument structure, they suffer from indeterminacy within languages and inconsistency across languages. Predicate composition accounts are both limited in the range of phenomena they tackle and fail to explain the full range of usage of the phenomena that they claim to address.

The approach of Dowty (1991) seems most promising, but it remains an incomplete theory over fifteen years after first being introduced. Many have worked at integrating its intuitions into various grammatical frameworks, but the scope of information that contributes to argument structure realisation remains largely unchanged. Dowty's account introduces flexibility into argument structure realisation, for example accounting for the cases in which *please/like* sentences and *load/swarm* diatheses are truth-conditionally synonymous, by making both variants available. But as a result he cannot account for cases in which one variant is preferred over another. Neither can he account for structures that seem to invert his prototypical relations, such as the passive construction, or verbs like *receive* or *undergo*. Discourse related features such as givenness, constituent weight, and anthropocentric saliency offer the promise to fill in some of these gaps.

An ideal theory of argument structure realisation would achieve several things. Firstly it should account for diathesis production, whether there is free variation among candidate realisations, one variant is preferred for discourse reasons, or one variant is sometimes necessary to express a crucial distinction in meaning. The theory would have to explain verb lexicalisation patterns, including atypical verbs such as *receive* (though they are small in number). And it should supply a measure of the relative wellformedness (or acceptability) of different realisation possibilities. Finally, the theory should be built on a basis that at least provides promise of a cross-language account of argument structure realisations.

Chapter 3

Empirical Investigations of Argument Structure Realisation

Linguistics as an enterprise has been criticised for the lack of rigour in its use of empirical evidence (see e.g. Robinson, 1975; Schütze, 1996; Cowart, 1997; Edelman and Christiansen, 2003; Wasow and Arnold, 2005). Elaborate theoretical conclusions are often drawn on the basis of small numbers of constructed examples, and judged by authors themselves.

In this work a primarily evidential approach is taken, which should prove more robust. Rather than taking an intricate model of processing or representation for granted, such assumptions are avoided to see what patterns emerge from the data – between argument structure and the syntactic, semantic and pragmatic features that have been suggested by the literature. In particular, an attempt is made to see if general principles of linking can account for phenomena that have previously been attributed to particular lexical or constructional entries.

The first set of experiments described (section 3.2) ask large numbers of native-speakers of German, English and Chinese to judge the acceptability of diathesis variants. Diatheses are of particular interest, since they constitute minimal pairs that manipulate argument structure while remaining lexically identical (excepting functional morphemes). In this way, differences in acceptability are due to diathesis manipulations only, as extraneous factors such as plausibility, elegance of expression and complexity of vocabulary are held constant. Here, a range of semantic and discourse factors are considered as possible determinants of diathesis success, to provide empirical support or challenge to theoretical accounts of argument linking (as discussed in the previous chapter). Between 25 and 50 carefully selected authentic examples of each diathesis construction are used. Large groups of participants are recruited via email, to complete the experiment over the web.[1] The materials are taken from popular cinema, to provide a familiar and accessible context, and all items presented for acceptability judgements are supplied with both background descriptions and surrounding text, to control interpretation by participants.

The diatheses chosen for the experiments all involve changes to the linking of nuclear arguments. The English dative/benefactive diathesis (section 3.2.2) allows a (usually) animate participant to appear either as a *for* or *to* prepositional phrase, or as a bare noun phrase indirect object of a verb in certain semantic classes. The English and German passive diatheses (sections 3.2.3 and 3.2.4)

[1] While web-administered experiments do not give the experimenter the control that a lab environment would provide, they permit the recruitment of a larger range of participants than would typically be available at a single physical location, at little or no cost.

have many structural similarities to each other, expressing the active canonical object participant as subject, and optionally realising the canonical subject as the object of a *by/von* oblique phrase. Functional accounts of the passive ascribe to it a perspective taking function that changes the focus over participants, and may draw attention to the process or result of the event being described. The Chinese *ba* and *bei* constructions (section 3.2.5) have many similarities to the English and German passives, involving the foregrounding of a canonical object, and a particular association with transitive events. In addition, two methodological questions of such acceptability experiments are investigated – the effect of background knowledge and textual context (section 3.2.6); and the choice of a rating scale for such judgements (section 3.2.7). The results of these experiments suggest that a range of semantic and pragmatic factors play a role in the selection of argument structures, and that these factors, many of which are active in more than one of the languages considered, have similar consequences for perceived acceptability to violations of syntactic constraints.

A follow-on corpus study tries to verify some of the conclusions of these acceptability experiments on a much larger scale (section 3.3). Instances of verbal arguments are extracted from fifteen million words of English text, and associations are examined between argument positions (canonical, grammatical and serial) and textually recoverable correlates of the semantic and discourse features of interest (e.g. animacy for agency; definiteness for givenness). This gives qualified support to the results of the acceptability experiments, though the effects seen are rather weak.

An annotation study (described in section 3.4) asks expert informants (linguists) to categorise the arguments of a set of arbitrary English sentences using a conventional set of discrete semantic roles. The differing levels of agreement found among judges can suggest which role categories are more or less successfully defined in the literature, and which roles may be candidates for combination into sets of generalised roles, such as those proposed by Dowty (1991), van Valin and LaPolla (1997) and Primus (1999). An extrapolation of the data gathered suggests that such reduced role inventories would lead to much higher level of consensus among judges (normalised relative to the number of role categories).

Before describing the experiments in detail, the chapter starts with a review of relevant methodology (section 3.1) in the selection of linguistic materials, the design of experiments that elicit judgements from informants, and considerations for the automatic analysis of corpus texts .

3.1 Notes on Methodology

This section discusses methodological choices that come to bear on the experiments described in this chapter. Selection of materials (section 3.1.1) is a necessary step in all the studies described. The diathesis acceptability experiments (section 3.2) and role annotation exercise (section 3.4) are both judgement tasks (3.1.2), the unsystematic use of which has led to challenges (Wasow and Arnold, 2005). Section 3.1.3 discusses considerations for corpus-based studies like that described in section 3.3.

3.1.1 Selection of Materials

Linguistic materials should be selected with several things in mind. Statistical methods demand that the data at hand is a random and representative (i.e. unbiased) sample of the phenomenon of interest, in order to be able to reliably estimate the strength of any conclusions drawn (Cowart, 1997, pp.119-122). The aim is that extraneous factors that the investigator is unaware of, or cannot

control for, may be factored out, at least in part. This also guards against unconscious bias on the part of the investigator to select items that best exemplify their theoretical viewpoint. In addition, experimenters must be conscious of the dialect, register and domain to which materials belong. If these factors cannot be balanced, then it should be made clear that any conclusions drawn are limited to the kind of material in question.

In corpus studies the experimenter must ensure that the text collection used is representative of the language (or variety) of interest. For such purposes, balanced corpora such as the British National Corpus or Susanne corpus are much superior to domain-specific corpora such as the Brown corpus (on whose basis the Penn Treebank was constructed). For corpus studies the methodological perils are rather in ensuring the adequacy of the automatic analyses performed (see section 3.3).

In smaller scale judgement experiments the investigator must typically search for appropriate items manually (though automatic methods can be used to aid the process). Ideally, materials should be authentic (i.e. attested) examples of language use, rather than constructed "toy" sentences. Sentences should be plausible, natural, use common vocabulary, and be unambiguous and easy to parse (Schütze, 1996, pp.31,184-186; Cowart, 1997, p.46). Neutral materials should be preferred, to avoid offence or particular personal resonance to participants (Cowart, 1997, p.46) – for example materials dealing with inter-faith relations might prove overly emotive for participants in Northern Ireland, Lebanon or Sri Lanka, and distort responses. For similar reasons, particularly evocative, iconic or amusing items may introduce confounds, but materials should not be so homogenous that they lead to boredom and lack of attention.

3.1.2 Judgement Experiments

Intuitive judgements of sentence acceptability and categorial membership are widely used in the linguistic literature to support theoretical conclusions. Often a small number of informants, including the investigators and their colleagues, are used to supply these judgements. As a result these judgements are open to bias, as participants' responses can be affected by repeated exposure to a particular sentence or class of utterances, and the inclination to view them in a way that supports their theory. In addition, some studies have shown language "experts" (e.g. linguists, editors, authors, language teachers) to have systematically different intuitions about acceptability to other speakers. Spencer (1973) found that a group of non-linguists agreed among themselves on the acceptability of 80% of a set 150 sentences drawn from linguistic literature. They agreed with linguists only 50% of the time. Even when experimenter bias plays no role, individual judgements can be highly unreliable, both across participants, and when a single participant is presented with the same item on more than one occasion (Cowart, 1997, pp.3-7,31-37).

This common practice of eliciting judgements informally has led many to have doubts about the empirical basis of linguistic theory (see e.g. Schütze, 1996; Wasow and Arnold, 2005). One response to this problem is to abandon judgements completely, and use objective methodologies such as EEG (electroencephalography – recording electrical activity in the brain by means of electrodes attached to the scalp), timed reactions and eye-tracking methods (described in more detail in section 1.3). While these methodologies have similar problems with inter-respondent differences, given an appropriate experimental design they can eliminate subjectivity from results: the fact that acceptability is being tested can be obscured (e.g. by asking participants to judge the truth of experimental items), or analysis can be performed on data recorded before conscious control of response is thought to begin (for example, some ERP components that are sensitive to linguistic violations appear within

the first 200 milliseconds of a sentence being presented). But such methods present considerable difficulties in interpretation. For example radically illformed sentences are similar to wellformed sentences in the time required to arrive at a judgement of acceptability – it is rather marginal or ambiguous sentences that require longer (Schütze, 1996, p.84). More generally, if a measurable difference is found between stimuli with any of these objective methods, there is not necessarily a clear indication of what factors are responsible for the difference. In such a case experimenters often make educated guesses (what Wasow and Arnold, 2005 refer to as "secondary intuitions").

Given the drawbacks associated with the alternatives, and the fact that acceptability experiments do access internal grammars (Schütze, 1996, p.180), even if in an unusual way (Schütze, 2005), judged acceptability experiments can be appropriate tools, if certain precautions are taken in the selection of participants, design of the task (including instructions), presentation of materials and analytical treatment of responses.

3.1.2.1 Selecting Participants

Individual judgements may not be reliable, varying both between participants (due to differences of idiolect), and when the task is given to a single participant on more than one occasion. However, aggregate judgements of acceptability can be consistent (i.e. are replicable) across groups of speakers if they are large enough, and of similar composition (e.g. all natives of a single dialect and non-experts) (Cowart, 1997, pp.80-83). The numbers required can be quite large – Cowart describes a study that needed 40 respondents to detect the difference in acceptability between the two sentences below:[2]

(3.1) a. Who did she sell the portrait of?

b. Who did she sell Max's portrait of?

Other than idiolect or dialect, possible confounds include age, handedness, gender and level of education (Schütze, 1996, pp.184-186, Cowart, 1997, p.46). Since it is often impractical to recruit groups that are either homogenous, or representative, of all these factors, participants can be selected at random. If participants are forthcoming, these details should be recorded so that they can be eliminated (or otherwise) at the data analysis stage. An effort should also be made to include participants that are not university undergraduates, as they represent a rather uniformly young, middle-class and well educated group, who are frequent readers.

3.1.2.2 Experimental Procedures

Any experiment should be first piloted with a small number of participants, with which the investigator can discuss the format of the experiment. However, in genuine experiments, the investigator should not take part, as they could be unconsciously providing cues to participants, and this also increases the pressure on participants to "perform". All of the experiments described in this thesis involved the written rather than spoken medium, for practical reasons – written materials and responses are easier to present and process, and can be delivered on paper or via software. Though the comprehension faculty is thought to be common to text and speech (see e.g. Eysenck and Keane, 1990, p.111), slightly different results might have been gathered if participants were responding to spoken stimuli (Murphy, 1997; Cowart, 1997, p.64). The internet provides a particularly easy

[2] Cohen, 1969, pp.60-64: the power of the experiment (i.e. the chance of detecting this difference, given the volume and composition of the response data) was 80%.

way both to recruit participants and to administer an experiment at many locations, though it gives little control over the conditions under which the experiment is taken – for example whether the participant is distracted or becoming bored (Schütze, 1996, pp.184-186). Another issue is that as the circle of respondents increases in size and distance from the experimenter, the variability in data can increase. In the judgement experiments described here, there was a noticeable increase in noise and a drop in the time spent per item in the sessions of web respondents, relative to acquaintances.[3] Presumably participants are more likely to be conscientious if they have some connection to the experimenter.

Web surveys have been shown to provide similar results to postal surveys, but with quicker and larger response rates. Wygant and Lindorf (1999) describe a comparison in which the same unpaid survey had a response rate of 50% when administered via the web, compared to 32% by post; the great majority (80%) of internet responses were returned within two days, while the same proportion of postal responses took 22 days to arrive. While the quantitative results for both delivery methods were similar, web responses to open-ended questions were longer in terms of word-count.

Participants should always be given some practice or warm-up items, as the first few responses have been shown to be of much higher variability than subsequent ones (Cowart, 1997, p.90-92). The warm-up items should be representative in terms of acceptability of the materials as a whole. This also provides a method to verify that participants have understood the marking scheme and are applying it correctly. For example, if responses to wellformed practice sentences are as a whole lower than those to the illformed practice sentences, it may be that the participant has inverted the scale.

All participants should be presented with the same or equivalent material. In studies where the materials are too much to reasonably present to a single participant, a random but representative subset must be selected, ensuring that the proportional composition remains approximately the same (e.g. all participants should see roughly similar proportions of test items to fillers and of acceptable items to unacceptable). Once a selection of items has been made for a participant, the order of presentation should be randomised also, but again items of each type should be spread as evenly as possible throughout the experiment (e.g. participants should not encounter all the acceptable sentences at the start and all the unacceptable items at the end), as a participant can become habituated, and so insensitive to distinctions (Cowart, 1997, p.46). Neighbouring experimental items can affect responses, but randomisation of composition and order controls both for this and the fact that people may become less careful as the experiment progresses (Greenbaum, 1976).

Test items should be interleaved with similar filler items, to prevent participants from discerning the purpose behind an experiment (Wasow and Arnold, 2005, p.1484), and from developing strategies for making judgements (Cowart, 1997, p.51-52) (e.g. a participant presented with equal numbers of passive and active sentences in isolation, might adopt a scheme of systematically judging passives below actives). Authors differ on what is the ideal proportion of test items to filler items, and of "good" items to "bad". Schütze (1996, p.185) suggests that there should be similar numbers of "good" and "bad" items, while Cowart (1997, p.91) advocates the use of two to four times as many filler sentences as test sentences. This presents the experimenter with the task of assembling large numbers of both acceptable and unacceptable filler sentences, which can be problematic. Au-

[3] In the English-language study described in sections 3.2.2 and 3.2.3, the first 'acquaintances' group of 48 participants spent a median 24 seconds per item, and the median interquartile range of responses was 0.58. The comparable 'strangers' group of 37 participants spent a median time of 20 seconds per item, and their median interquartile range was 0.93. Both effects were significant at $p < 0.05$ (two-way independent samples t-test over participant sessions with Welch correction for unequal variances). Similar patterns were found in the German and Chinese studies, but they were weaker and not statistically significant.

thentic unacceptable sentences are hard to come by, and the particular set of "bad" fillers selected will affect the responses to the experiment as a whole, reducing the likelihood of other researchers being able to replicate results (without using precisely the same materials). In this study only fully authentic sentences are used as fillers.

Items should ideally be presented together with a supporting context, as the task of acceptability judgement necessarily involves assigning a meaning, and the experimenter cannot know what a participant's imagined interpretation of isolated sentences is (Wasow and Arnold, 2005, p.1485). Section 3.2.6 will demonstrate to what extent explicit context (where the situation context is provided) and implicit context (where it is already known to the participant) increases the reliability of responses, and so improves statistical robustness.

Instructions play a very important part (Cowart, 1997, p.56-58), as some early experiments in acceptability showed: Schütze (2005, p.458) reports one study in which a participant uniformly marked imperatives down, because they did not include the word *please*. One cannot be sure to what extent the responses that participants report are purely intuitive, impressionistic and unconscious, or are based on rules (whether prescriptive or an ad-hoc schema the participant has come up with on the fly) (Schütze, 1996, pp.81-84). The notion of acceptability can be grounded in a situation that is easy to imagine, such as helping to edit a paper, or correcting the usage of a friend who is learning your language. Participants should be explicitly directed to follow their intuitions or gut-reactions, and not to base their responses on rules of grammar or punctuation that they may have been taught, and appropriate examples should be given for illustration. It is generally recommended that the notion of acceptability should be expressed in terms of naturalness (e.g. does an item sound normal or awkward) and ease of understanding, rather than "right" or "wrong", though Cowart (1997, p.57) reports an experiment on remote antecedents in which the choice of instructions varied from a naturalistic approach ("fully normal" versus "very odd, or awkward") to a prescriptive one ("completely grammatical" versus "not regarded grammatical by an appropriately trained person") without a substantial effect on results.

3.1.2.3 Judgement Scales for Acceptability

Since acceptability does not appear to be an absolute binary notion (see section 1.3), some sort of relativised scale must be used to elicit judgements. Likert scales are commonly used to measure attitudes in surveys and judgement experiments, and use a fixed number of ordered categories. The categories can be phrasal (e.g. "strongly agree", "agree", "no opinion", "disagree" and "strongly disagree"), numerical (e.g. +2, +1, 0, -1, -2), or numerical with the extremities and zero point labelled.[4] Many authors have pointed out that one cannot assume that the intervals between points on a Likert scale are of equal magnitude, though researchers often do so when taking the arithmetic mean of such data (Jamieson, 2004). Surveys commonly use five point Likert scales, but the fixed number of categories will necessarily distort responses. Participants may find that at some point in an experiment they have already used the most extreme values on the scale, and when they encounter an item that is yet higher or lower on the scale than all that preceded it, they are not able to signal the distinction. Similarly, the rounding involved in fixed categories can distort the distances between items: for example, given three items that would have a numerical score of 5.6, 6.4 and 6.6 on a

[4] Strictly a Likert scale involves numbers appearing alongside phrasal labels as ratings are made. Some of the work here adheres to this exactly, while some supply labels at the instruction stage but only numbers during the experiment. Such categorial scales have a long history. Lodge (1981, p.5) mentions a six-point brightness scale used to describe stars in Greece in 150 BC.

finer-grained scale, rounding will reduce them to 6, 6 and 7, falsely indicating that the first two are more alike than the second two. Further, if the acceptability of an item corresponds to a rating that lies directly between two available categories, participants must arbitrarily choose one or the other, introducing unnecessary variation in responses (so increasing noise and reducing the sensitivity of statistical tests). A final problem is that of relative scaling between participants. Though the addition of phrasal labels can mitigate this somewhat, one cannot be sure that participants interpret a scale in the same way – a score of +2 may signify a higher value for participant *A* than it does for participant *B*.

Other methods of arriving at a ranking of experimental items (Schütze, 1996, pp.78-81) include set ranking. Participants can be presented with a list of items and asked to order them, though there is clearly a limit to how many can be reasonably compared at once.

Items can also be presented for pairwise comparison, where each item is compared to many others (ideally to all others). Items can then be ranked on the basis of the number of items they were rated above. If large numbers of participants can be found, random pairwise comparison can be appropriate. With each item being compared to an arbitrary but large subset of the experimental items (i.e. by many participants), a percentile score can still give a good aggregate estimation of acceptability, and these scores constitute an interval scale (relative to the set of experimental materials). Drawbacks include the number of responses are needed, and the fact that pairwise comparison responses do not always prove to be transitive – participants may rate item *A* above *B*, *B* above *C*, but *C* above *A*.

Magnitude Estimation (Stevens, 1975) is an adaptable scaling technique from the field of psychophysics: establishing the correspondence between physical stimuli (e.g. heat, colour, loudness) and perceived intensity. Participants are asked to consider a reference stimulus and give it an arbitrary rating, and then rate each subsequent stimulus relative to the reference. The technique is explicitly ratio based, in that participants are asked to rate each stimulus by how many times larger or smaller it is relative to the reference. There is no maximum or minimum rating to constrain participants, and gradations between items can be as fine as is required. Stevens formulated the Psychophysical Power Law (1975, pp.13-19) (subsequently experimentally confirmed for a range of physical continua), that states that there is a linear relationship between the logs of the stimulus ϕ and response ψ magnitudes. The exponent β describes the slope of the line, and is characteristic both of the physical stimulus in question and of the unit of measurement used.

(3.2) $\psi = k\phi^{\beta}$, equivalent to $\log\psi = \log k + \beta\log\phi$

Participants usually express the rating by assigning a numerical value (in a laboratory environment other modalities can be used, such as drawn line length). This law was shown to hold also when the stimulus/response pair were reversed, for example when a participant was asked to use an unmarked volume control to set a loudness level corresponding to a numerical stimulus (termed magnitude generation).

For analysis, magnitude estimation response scores are log normalised. If a base of two is used, the responses have an intuitive interpretation: a zero score signifies equality with the reference stimulus; a positive score of one signifies a magnitude of twice the reference, and negative one half the reference; positive or negative two signifies four times or a quarter of the reference respectively. A central tendency in the responses from multiple participants to a single item is estimated by the arithmetic mean (in terms of ranking of estimates, this is equivalent to taking a geometric mean of the raw unlogged values; Stevens, 1975, p.263).

Early applications of magnitude estimation (or magnitude scaling as it is also known) to non-physical domains included a study of public attitudes to the Swedish monarchs and an investigation of perceived word frequency (Stevens, 1975, pp.238,263). For attitudinal phenomena, where there is no physical continuum with which to establish the power law given above, experimenters developed a method of cross validation to confirm that the power relationship held between submitted scores and the internal psychological scales associated with opinions. To do this a single attitudinal scale can be measured against two external response scales. Once the power law relationship is established for both stimulus-response pairs, these slopes (that is the exponent β for each) can be used to predict the power law relationship between the two response scales. The first such published experiment rated attitudes on racism using both loudness and hand grip pressure. The power law function between hand pressure and loudness correlated with the follow-on experiment very highly ($r = 0.97$, Lodge, 1981, p.32).

Bard et al. (1996) applied the same cross validation technique to judgements of linguistic acceptability. They tested acceptability judgements of the Italian *avere* and *essere* auxiliary constructions by native speakers using the magnitude estimation technique in two modalities: number assignment and line length (they recommend that even in live experiments, a practice session should be carried out against a physical quantity such as brightness to familiarise participants with the technique). A very high correlation of $r = 0.99$ was found between modalities.

Keller (2003) examines a series of experiments that use magnitude estimation with acceptability judgements of various phenomena in several languages. He characterises each linguistic phenomenon in terms of constraints that contribute to acceptability, and proposes to use the number of constraints violated as the "physical" scale to which judgement data might be related by Stevens' power law. A power law model provided good correlation between number of constraints broken and acceptability judgements (in seven experiments, correlations ranged $0.81 \leq r \leq 0.93$), noticeably better than with a linear model ($0.72 \leq r \leq 0.88$). However the characteristic slope of each power function varied dramatically. Three separate experiments on Greek word order variously yielded an exponent β of 0.14, 0.83 and 0.30, though Keller notes that introducing weightings to the constraints might provide a more consistent conclusion.

However, there are sources of distortion inherent in this methodology. Even in ideal experimental contexts, the relationship between stimulus and response does not follow the law perfectly. Participants show a slight conservative tendency, underestimating very extreme stimuli. Conversely, participants tend to make full use of the scale, and will stretch out the range of their responses if all the stimuli are within a small range (Stevens, 1975, pp.281-286). When the response scale is numerical, responses are further distorted by the tendency for participants to use round numbers, and to avoid crossing certain arbitrary limits, such as using decimals below one, or scores of more than one hundred. In the context of software based testing, using line length is not an option, due to the space constraints of computer screens. Numeracy can be an issue, as not all participants may be comfortable or reliable when multiplying and dividing numbers (Cowart, 1997, p.73) – and if participants use rough-and-ready estimation to avoid difficult arithmetic, that will introduce noise into the data. But despite these caveats, magnitude estimation is claimed to result in less noisy data than similar categorial techniques. Lodge (1981, p.79) mentions a study in which data elicited using magnitude estimation explained 12 to 15% more of the variance in the data than a categorial scheme.

The choice of judgement scale also has consequences for the methods of analysis that are appropriate. Stevens (1946) introduced a widely used four-way typology of data types (nominal, ordinal, interval and ratio data), together with a controversial proscription of what statistical tests are appro-

priate to each (see e.g. Michell, 1986; Velleman and Wilkinson, 1993). Steven's point as regards numerical data is that the arithmetic mean presumes that the difference between subsequent points on the scale are an equal distance from each-other. This cannot be said of ranked (ordinal) data, and so parametric tests that are based around means (such as the *t*-test) are unsafe. While some textbooks (e.g. Grimm, 1993) adhere to this classical model, others avoid a fixed data typology completely (Moore and McCabe, 2003), and emphasise instead the importance of sample size in deciding the appropriateness of statistical tests. For Moore and McCabe (2003, pp.187f) arbitrary transformations (such as a log function) are justified if they reveal an interesting relationship among the data and produce scientifically meaningful results. Magnitude estimation is said to permit the use of the more sensitive parametric statistical techniques by virtue of its yielding interval data, though this does not free it from the other demands of common parametric tests: that the data have a Gaussian sampling distribution; and roughly equal sample sizes and variances, when comparing more than one mean.[5]

In his typology Stevens makes a broad assumption about the nature of scientific data – that formal mathematical quantities can directly represent empirical facts. Alternatives are to adopt either an operational or classical stance (Michell, 1986), when examining non-physical phenomena. The operational stance views measurements as phenomena in themselves, and conclusions are not made about any underlying reality (which approaches behaviourism in the context of psychology). The classical stance sees measurement scales as the best way to access fundamentally inaccessible variables.

The choice of marking schemes is thus a compromise between conflicting goals of data resolution and statistical robustness, against ease of understanding and use for participants. Too simple a system does not allow participants to express subtle differences and can compress real differences in acceptability; but too complex a system may lead to confusion and introduce errors into the data. Many participants find the magnitude estimation task "bizarre" initially (Bard et al., 1996, p.41), and this lack of face validity may discourage respondents in cases where motivation is not high (for example, the unpaid web respondents used in these studies).

The use of magnitude estimation for judgements of linguistic acceptability has both its supporters (e.g. Cowart, 1997, pp.70-74) and detractors (e.g. Schütze, 1996, p.79). Magnitude estimation was chosen as the primary measurement scale used in the acceptability experiments described in this thesis. Its relative utility in obtaining experimentally robust acceptability data is compared both to a Likert scale and to pairwise comparison in section 3.2.7.

3.1.2.4 Annotation and Categorisation Exercises

Categorisation tasks involve assigning an item to one of several non-ordered (i.e. nominal) predefined classes. A common categorisation task in linguistics is the annotation of text with syntactic, semantic or discourse categories. Judges must already be acquainted with the categories in question, or be supplied with extensive guidelines, or training in applying them.

Several statistics of agreement can be used to measure to what extent different judges agree or disagree on categorisations. A simple unnormalised measure of intra-item agreement is the plural-

[5]Different statistical tests are more or less robust against different violations. The Gaussian assumption is satisfied to different degrees, depending on the size of samples examined (cf. LaPlace's Theorem). The question of equality of variances can be addressed by using modifications of common tests (such as the Welch adjustment to the *t*-test). The published literature on the use of magnitude estimation in linguistics does not address these assumptions of distribution and variance. This is a pertinent question, since the open-ended nature of the magnitude estimation scale makes it (pre-theoretically) more prone to anormal distribution and different variances than a bounded scale such as a Likert scale.

ity vote (that is the proportion of responses for that item that chose the mode, or most common, category). Alternatively, the proportion of agreement can be calculated as the number of pairwise agreements found between k judgements, divided by the number of pairs of judges (k choose 2). Such indices of agreement, that vary between 0 (no agreement) and 1 (total agreement), are often normalised relative to the expected proportions of agreement. In the absence of prior knowledge of the general distribution of the categories at hand (that is assuming all categories are equally probable) the expected proportion of agreement *P(e)* is the inverse of the number of categories, $1/n$. The pairwise agreement measure a_{p} for a *single item* (3.3) yields 1 for total agreement (all judges agree on a single categorisation), and 0 if agreement is exactly that expected by chance. A negative value indicates that less agreement was found than chance would predict.

(3.3) $a_p = \frac{P(a)-P(e)}{1-P(e)}$, *P(a)* is proportion of observed agreement, *P(e)* is proportion of expected agreement

A measure of agreement for a complete *set of items* can also be calculated on a similar basis, first by averaging the by-item agreement proportions. Expected agreement is then estimated on the basis of the frequency with which each category C_{j} was used across all items – the probability p_{j} of selecting such a category is this count divided by the total number of judgements made (3.5). If pairwise agreement between judgements is the measure of interest, the resulting statistic is known as kappa K (see Siegel and Castellan, 1988, pp.284-291; Carletta, 1996):[6]

(3.4) $P(A) = \frac{\Sigma P(a)_i}{m}$, for m items

(3.5) $P(E) = \Sigma p_j^2,\ p_j = \frac{C_j}{N}$, where C_{j} is the number of times a single category was chosen, and N is total number of judgements made

(3.6) $K = \frac{P(A)-P(E)}{1-P(E)}$

For plurality measures of consensus, expected agreement is simply the relative frequency of the most commonly chosen category (assuming all categories are equally spread across items, the most common category for each item would be the same).

(3.7) $P(E) = p_{max}$, where p_{max} is the probability of the most frequently annotated category

The difference between these two measures is that the K statistic counts *all* agreement between judges, while the plurality based measure only counts agreement with the correct annotation, according to that set of judges. Both measures of individual and global agreement will be used in the annotation study in section 3.4.

3.1.3 Corpus Experiments

Due to the work involved in assembling appropriate materials and recruiting participants to take part, the resulting data sets from psychometric experiments are often limited in size. There are also deeper questions of the authenticity of the tasks presented to experimental participants, and to what extent they reflect the people's language capacity and usage. Corpus experiments can provide dramatically larger quantities of data. For example the portion of the British National Corpus of English used here contains over one million sentences and has been carefully assembled (by committee decision)

[6] Artstein and Poesio (2007, p.1) note that there are many terminological inconsistencies in the literature. What Siegel and Castellan term K is a "generalisation to more than two coders of Scott's π".

to give a balanced representation of language use across domains, registers and regions (see section 3.3.2). The world wide web also provides a massive and freely available repository of raw text, though questions of how representative it is, and about the reliability of frequencies returned by search engines remain (Kilgarriff, 2007).

The judgement experiments performed here rarely dealt with more than one hundred items at a time, and usually used data from a single type of text (e.g. film scripts, web pages). The advantages of corpus data in terms of volume are offset somewhat by several weaknesses to this approach. Firstly, corpora only constitute positive evidence – they tell you what people do say and write, not what they don't. Secondly, though automatic parsing methods have made huge progress over recent years, they remain less reliable than humans on many tasks. Thirdly, the reliability of contemporary automatic methods with structural properties of a text do not extend to semantic and pragmatic distinctions (for example current word sense disambiguation systems struggle to achieve 70% accuracy, while simply choosing the most frequent sense of a word can yield 60%; Mihalcea and Csomai, 2005). State of the art parsers such Miyao and Tsujii (2005) or Burke et al. (2004) (variously described as deep syntactic or shallow semantic parsers), are limited to identifying grammatical function (verb argument positions).

However, when looking for patterns in large numbers of sentences, it may be possible to infer semantic and pragmatic features from structural features. For example, definiteness of nominals is a good indicator of the givenness in discourse of the entity they refer to (see e.g. Chen Ping, 2004). Similarly the animacy of a noun phrase is usually a necessary (but not sufficient) condition for its playing the role of agent or experiencer in a sentence. This is the approach adopted here – to supplement the relatively reliable, but small scale results of human judgements with the larger scale, but rougher grained results of corpus investigations.

3.1.3.1 Tagging and Parsing

While part-of-speech taggers typically achieve full coverage, and accuracy measures in the high nineties (e.g. 97% for an early statistical tagger, Brill, 1992), sentence parsers are less reliable. A state-of-the-art statistically based tree parser (e.g. Collins, 1999; Charniak, 2000) typically achieves around 90% accuracy (0.9 in terms of the conventional measure of F-score[7]) per node of constituent structure, on restricted domain text (English language statistical parsers are almost exclusively trained and evaluated on a single medium sized corpus of newspaper text, the Penn Treebank; Marcus et al., 1994). In this study Minipar – a rule-based and lexically-driven dependency parser – was used. A comparison (Lin Dekang, 1998) of Minipar output with the hand-parsed Susanne corpus (Sampson, 1995) achieved precision (i.e. correctness) of 88-89% and recall (i.e. coverage) of 71-79%,[8] over all subject and complement relations (see section 3.3.3.1 for more discussion).

[7] Statistical parsers are conventionally evaluated by comparing an automatic parse against a "gold-standard" (i.e. benchmark) manual parse. The accuracy of the parser on a single sentence is calculated over nodes in its syntax tree, and is then averaged over the whole test corpus. Correctness (p "precision") is the proportion of the automatically parsed nodes that are also found in the manual parse, and coverage (r "recall") is the proportion of manual parse nodes to be found in the automatic parse. These figures are combined in the F-score (which originated in information retrieval) as follows: $F = 2pr/(p+r)$. When p and r are equal, $F = p = r$. An F-score of 0.9 indicates that both accuracy and coverage lie between 0.8 and 1.0.

[8] Equivalent to an F-score of $0.79 - 0.83$.

3.2 Diathesis Acceptability Experiments

3.2.1 Introduction

As discussed in the introduction and review chapters, some accounts have claimed that the productivity of certain diatheses is licensed exclusively at the syntactic (e.g. Chomsky, 1995) or lexical (Rappaport and Levin, 1988) level. However certain experimental results (e.g. Tanenhaus et al., 1995; Hagoort, 2003) cast doubt on such modularity, suggesting that elements of discourse and of interpretation are integral to the language processing faculty. In this section the effect of various factors are investigated, to determine if elements of semantics and pragmatics, both enduring and context dependent, have consequences for the acceptability of diathesis forms, and if such violations are of similar magnitude to those seen when syntactic rules are broken.

The diatheses to be examined are all reasonably common in use, have been well-studied in the literature, and involve manipulations of nuclear grammatical functions: the English dative/benefactive diathesis; the English and German passives; and the Chinese *bei* and *ba* constructions.

For each of the diatheses, authentic test sentences were gathered that were structurally suited to the manipulation in question (active sentences for passive, *bei* and *ba* constructions; Verb–NounPhrase–PrepositionalPhrase sentences for the dative/benefactive). For each a diathesis variant was also manually constructed. These materials were then presented in randomised order, and interleaved with filler material, to many participants, both linguists and non-linguists, who were asked to rate them for wellformedness. No participant saw both diathesis variants of the same experimental item.

Film scripts were used as the source, as they constitute a domain that is familiar to the general participant, and so that experimental items could be presented with textual context (surrounding dialogue) and background descriptions, so controlling interpretation. This approach has some draw backs, in that particularly evocative or entertaining dialogue excerpts may elicit higher judgements. It may also be that participants who are cinema enthusiasts may remember scenes word-for-word, and give lower judgements to altered sentences that are nevertheless well-formed. However, it was felt that despite such issues, film scene materials would be superior to using constructed examples, or authentic text that is obscure, unfamiliar or uninteresting.

Each of the diatheses in question consists of an unmarked/marked pair. The unmarked is the more general construction, which can be replaced with the marked construction, subject to some syntactic, semantic and pragmatic restrictions – the nature of which form the object of these experiments. For the English and German active/passive diatheses and the Chinese active/*ba* and active/*bei* diatheses, the active sentence is the unmarked one. For the English dative and benefactive diatheses, the Verb-NounPhrase-PrepositionalPhrase patterns is taken as the unmarked variant, as this is more generally applicable – restrictions on usage related to animacy operate on the double NounPhrase variant. This view is also supported by the fact that indirect objects are syntactically peculiar. e.g. in nominalisations, compare direct object-nominalisation compounds *secret-telling*, *book-reading*, with those using indirect objects, **spy-telling, *child-reading* (Baker, 1997, pp.13-20).

As was mentioned in the introduction (1.2), these experiments assume that the diatheses in question preserve literal meaning. The experiment by Honeck (1973) suggested that this is the case for the passive construction (see page 10). In the second chapter (section 2.5) it was argued that this is also the case for the dative diathesis. Divergent interpretation of diathesis variants is also constrained by the accompanying textual context.

Certain theories of argument linking can be challenged with the acceptability judgements given

to single examples. For example, if a non-metaphorical use of the dative double-object construction with a non-recipient indirect object can be shown to be acceptable to native English speakers, this poses problems for a theory such as Pinker's (1989). The influence of single semantic and discourse features will be established by showing statistically robust effects on the success of diathesis variants, relative to their unmarked counterpart.

Since diathesis variants constitute minimal pairs (to the extent that they share the same vocabulary and meaning), any differences in acceptability between them can be ascribed to constraints on argument linking. Each experiment has a principle hypothesis that the animacy of object constituents has a determining effect on the success of the diathesis, which is quantified as the differential judged acceptability between authentic original variants and manually altered variants. Posterior hypotheses are also tested on the effect of a wide range of features of semantics and discourse on diathesis success.

This section starts with individual presentations of the diatheses in general terms (3.2.1.1-3.2.1.4). A description of the process of manually coding experimental materials for animacy and other semantic and discourse features follows (3.2.1.5). The three language experiments are described in sections 3.2.2 to 3.2.5. Finally, the effect of methodological choices in the provision of context, and of choice of judgement scale are examined (sections 3.2.6 and 3.2.7).

A note on terminology is necessary here. Arguments will be referred to by the grammatical relations that they take in the unmarked "canonical" form – that is the active sentence for items in the Chinese *bei/ba,* English passive and German passive experiments, and the NounPhrase–PrepositionalPhrase configuration for the English dative/benefactive diatheses.

3.2.1.1 The English Dative and Benefactive Constructions

The dative and benefactive are often treated together, as both involve an unmarked NounPhrase–PrepositionalPhrase structure (the dative with *to*, and the benefactive with *for*) and a marked form in which a bare NounPhrase indirect object appears between the verb and the NounPhrase in the unmarked form (the direct object).[9] Semantically they are related also, in that being a recipient is often of benefit, and beneficiaries are often ultimate recipients, and for some verbs the double object construction is equivalent to a sentence using either *to* or *for* (3.8). However, recipients are not necessarily beneficiaries (3.9a), nor are beneficiaries necessarily recipients (3.9b).

(3.8) a. I'll bring you something back from my holidays

b. I'll bring something back to you from my holidays

c. I'll bring something back for you from my holidays

(3.9) a. I sent the letter bomb to Séan

b. I sent the letter bomb for Séan

The dative diathesis (3.10) only applies to verbs that involve a transfer between participants, whether concrete (e.g. *give, send*) or abstract (e.g. *teach, tell*). But this semantic class is subject to arbitrary exceptions: for example *donate* is semantically close to other verbs of giving but does not permit *he donated them his house*.[10]

[9] For example, Goldberg (p.75 1995) considers the benefactive usage as a sub-sense of the dative construction.

[10] Some have suggested that the condition is phonological or morphological in nature, and that Latinate verbs are excluded (Levin, 1993, pp.45-48)

(3.10) a. It's Maui Waui, I got my surfer-ex to **send me**$_{\text{iobj}}$ **some**$_{\text{dobj}}$.[11]

b. It's Maui Waui, I got my surfer-ex to **send some**$_{\text{dobj}}$ **to me**$_{\text{iobj}}$

The benefactive diathesis (3.11) applies more generally to verbs of creation (e.g. *build, bake, design*), performance (e.g. *sing*) and obtaining (e.g. *get, bring*, *steal*), but again there are exceptions such as *acquire*, *embezzle* and *construct*.

(3.11) a. I baked some currant buns for you$_{\text{iobj}}$.[12]

b. I baked you$_{\text{iobj}}$ some currant buns$_{\text{dobj}}$

Verbs that do admit the dative and benefactive diatheses are further limited in their application by features of their unmarked prepositional arguments (or alternatively by the nature of their relationship with those arguments). Some have stated the generalisation simply in terms of animacy, saying that animate prepositional arguments can appear in the post-verb indirect object position, and inanimate ones cannot.[13] This restriction only appears to apply to literal usages. Idiomatic, phrasal verb or abstract usages such as those below allow inanimate indirect objects:

(3.12) A House committee which heard his local option proposal is expected to **give it a favorable report** ...

(3.13) ... it seems logical to expect that polyphosphate ions would be strongly sorbed on the surface of the dirt (especially clay soils) so as to **give it a greatly increased negative charge**.

Some authors see subtle differences in meaning between dative and benefactive diathesis forms (e.g. Pinker, 1989; Krifka, 2004), but as already discussed in section 2.5 (see pages 50 on), attested counter-examples argue against such views. However, one or the other variant can be preferred for pragmatic purposes: to move a heavy constituent to the end of a sentence; or to give the constituent the particular focus and intonation associated with the final constituent (Quirk et al., 1985, pp.1208-1209; Hewings, 1999, p.84).[14] Various authors (Aissen, 2003; Bresnan and Nikitina, 2003; Wunderlich, 2006) have suggested that properties of information structure and saliency may license the dative diathesis.

3.2.1.2 The English Passive Construction

English does not generally allow free swapping of argument positions, as German and Chinese do.[15] The passive, which inverts the canonical subject/object order is the most common means of doing so – in written English, the passive occurs roughly once for every thousand words of text in mixed corpora (Biber et al., 1999, pp.939,961), or in about once in six sentences (e.g. in the Susanne

[11] BNC text A0L: Fiction, *Jay loves Lucy*, Fiona Cooper, 1991.

[12] BNC text CK9: Fiction, *The rag nymph*, Catherine Cookson, 1992. Repeated from page 6.

[13] The COMLEX computational English dictionary uses these patterns as an operational definition of what it views as a syntactic feature of animacy (Grisham et al., 1994) – any noun found in corpora to appear as the first noun phrase in a double noun phrase of known dative and benefactive verbs is listed as animate in this lexicon.

[14] While intonation is surely an integral component of argument structure (see e.g. Keller and Alexopoulou, 2001), the written materials used here do not allow it to form part of this analysis.

[15] Sometimes it is possible to move objects in front of the verb, but the subject must remain in situ:

(3.14) a. I$_{\text{subj}}$ hit James$_{\text{dobj}}$ really hard

b. James$_{\text{dobj}}$ I$_{\text{subj}}$ hit really hard

Corpus, Sampson, 1995). It is perhaps also one of the most well studied constructions in English - the LLBA[16] collection of linguistics abstracts lists 143 journal articles on the subject in the 5 last years alone. Many active sentences, such as (3.15a) have a passive formulation such as (3.15b) that is equivalent in meaning, as well as the underspecified (3.15c). In the English passive, the verb's canonical object appears as subject, the canonical subject is pushed to the end of the clause, or is omitted, and there is often a change of emphasis from the execution of the event to its result. The primary function of the construction is in discourse structure, allowing information to be emphasised or de-emphasised, flagged as new or given, or to give a better flow to the text.

(3.15) a. Big Brother expelled Jack Dee

b. Jack Dee was expelled by Big Brother

c. Jack Dee was expelled

Keenan (1985, pp243ff) gives this typological characterisation of the passive. Though not universal, *typically* in the world's languages, passives:

- Exhibit foregrounding and backgrounding of participants
- Are syntactically marked and have a different argument structure
- Are stative (i.e. show affectedness)
- Are formed at the verb phrase (VP) level, not the sentence level

With respect to the active sentence, *Jack Dee* is foregrounded to the prominent position of subject in (3.15b) and (3.15c), *Big Brother* falls back to the less prominent object position or disappears altogether, and the finite verb has been replaced by it's past participle, embedded under the verb *be*. *Jack Dee* is also clearly an affected party in the telic event described.[17] As no commitment is being made to any particular syntactic theory in this experiment, the last theory-specific point is not considered. But otherwise the English passive displays all of Keenan's characteristics.

There is also some variation in the 'passiveness' of sentences in English. Biber et al. (1999, p.936) differentiate between dynamic passives and stative passives. Stative passives can often be recognised by the fact that the participle shows characteristics of an adjective. For example in (3.17) the apparent lack of the finite verb *untrain* suggests that *untrained* is an adjective.[18] Stative passives can also be recognised by their ability to be coordinated with another unambiguous adjective (e.g. *naive*) and modified with a graded adverb (e.g. *relatively, rather, very, quite, more*) (3.17c). This seems to confirm that these constructions are more appropriately analysed as copula-adjectival constructions, rather than passives.

[16] Linguistics and Language Behavior Abstracts, http://www.csa.com/factsheets/llba-set-c.php; figures as of 26th July 2005

[17] Keenan's grouping together of stativity with affectedness can be criticised. A passive can be stative without expressing affectedness (e.g. *that is known by many*) or express affectedness in a dynamic way (*his shoes were being ripped to shreds by a Chiwawa*).

[18] An alternative view would be that *untrained* is indeed a finite verb or a verbal participle that appears in some form of a reduced clause, somewhat like a reduced relative. Distributional evidence does not support this view – consider the closely related verb *retrain*. The token *untrained* appears on approximately 872 thousand web pages in the Google internet directory, outnumbering *retrained* (210 thousand pages) by 4 to 1. If *untrain* were a fully-fledged verb one might expect a similar pattern of incidence of unambiguously verbal forms. On the contrary, *retrains* outnumbers *untrains* by 55 to 1, and *to retrain* is over 250 times more common than *to untrain*. A third view is that the rarity of *untrain* is not an indication of its absence as a verb, but rather that situations rarely arise to which it can apply. Perhaps a more conclusive case is that of the participle *destined*. In the BNC, *destined* appears 806 times, and never in the other verbal forms *destine, destines, destining*. The Oxford English Dictionary lists the transitive verb *destine* as obsolete, the latest attested appearance being in 1881:

(3.16) This is the place, O ye heavens! which I destine and select for bewailing the misfortune.

(3.17) a. She was untrained and, in one botched job killed a client.[19]

b. *Someone/they/people untrained her and, in one botched job she killed a client.

c. She was naive and relatively untrained, and in one botched job killed a client.

Quirk et al. (1985, pp.159-169) talk of the 'passive gradient', as demonstrated by the constructed examples they give from most central (3.18a) to most marginal (3.18i). They consider examples (3.18a-d) to qualify as true passives, identifying those with an agentive personal by-phrase as the most prototypical, while examples (3.18e-i) are termed pseudo-passives.

(3.18) a. This violin was made by my father

b. This conclusion is hardly justified by the results

c. Coal has been replaced by oil

d. This difficulty can be avoided in several ways

e. We are encouraged to go on with the project[20]

f. Leonard was interested in linguistics

g. The building is already demolished

h. The modern world is getting more highly industrialized and mechanized

i. My uncle was very tired

The separation can be justified by the existence (3.19a-d), or absence (3.19e-i), of a grammatical and semantically consistent active counterpart. Omitted canonical subjects can often be reconstructed with a an unspecific *someone*, *they* or *people* (Hewings, 1999, p.60), but as Quirk et al. (1985, pp.159) put it, the active-passive correspondence should proceed "without a change in the facts".

(3.19) a. My father made this violin

b. The results hardly justify this conclusion

c. Oil has replaced coal

d. We/They/... can avoid this difficulty in several ways

e. @Someone/They/People encouraged us to go on with the project.[21]

f. #Someone/They/People interested Leonard in linguistics

g. #Someone/They/People demolish(es) the building already

h. #Someone/They/People are industrializing and mechanizing the modern world more highly

i. *Someone/They/People very tired my uncle

A final test for stative-passives is the substitution of another copular verb for the verb *be* (3.21).[22]

[19] Penn Treebank 2.0

[20] Quirk et al. (1985) make it clear that they intend the stative reading of this sentence, rather than that meaning 'We are being encouraged to go on with the project'

[21] @ is used as a mark of semantic divergence in diathesis. For example the dynamic passive sentence (3.19e) does not mean the same thing as its stative counterpart (3.18e). See page 183 of the appendices for a list of notations used.

[22] However there is also a class of *get*-passives that show all the signs of a prototypical passive other than the choice of embedding verb (3.20), though they would be clear pseudo-passives for Quirk et al. (1985). This example comes from the Susanne Corpus 5.0:

(3.20) a. He got hit from the blind side by the split end coming back on the second play of the game.

b. The split end hit him from the blind side coming back on the second play of the game.

(3.21) The California Democrat is/appeared/remains/seems/got embarrassed by provisions inserted on behalf of owners of private beaches in the Virgin Islands.[23]

Syntactic, Semantic and Pragmatic Restrictions The primary structural requirement for verbs to appear in the passive is that they are transitive (Quirk et al., 1985, 3.68); i.e. they canonically take at least one bare noun-phrase object.[24] Idioms may (3.22) or may not (3.23) appear in the passive.

(3.22) a. Shawna delivered the goods

b. The goods were delivered by Shawna

(3.23) a. That does the trick

b. #The trick is done by that

Reflexive (3.24b), reciprocal (3.25b) and similar constructions with an anaphoric relationship between subject and argument (3.26b) can present problems for the passive.

(3.24) a. She carefully washed herself

b. ?Herself was carefully washed by her

c. ?She was carefully washed by herself

(3.25) a. They destroyed each other's reputations

b. ?Each other's reputations were destroyed by them

c. ?Their reputations were destroyed by each other

(3.26) a. She broke her leg

b. @ Her leg was broken by her

While some (e.g. Fillmore, 1968) have viewed the admissibility of the passive as lexically determined, a semantic restriction often noted is that verbs are typically dynamic. Stative relational verbs, tend to passivise clumsily, or not at all (3.27b)-(3.30b) (Quirk et al., 1985, 3.68; Palmer, 1994, p.119). There are exceptions (3.31b), and the same meaning can often be paraphrased with another copular structure (3.27c)-(3.28c).

(3.27) a. The novel lacks a strong plot-line

b. ?A strong plot-line is lacked by the novel

c. A strong plot-line is lacking in the novel

(3.28) a. I had that car for years

b. #That car was had by me for years[25]

c. That car has been mine for years

(3.29) a. The opera lasted/occupied/used up many hours

b. ?Many hours were lasted/occupied/used up by the opera

[23] Penn Treebank 2.0.

[24] Passive constructions of phrasal verbs and of prepositional objects are also possible but rare (Hewings, 1999, p.58). Clausal objects are possible, but not always successful (Quirk et al., 1985, 3.70), usually being alternatively realised as constructions of the form *it is known/thought/said that*.

[25] Archaic forms like *That car had I for years* may be possible, but pronoun case marking and the lack of *by* show that these are not passives.

(3.30) a. The ticket cost 20 euros

b. #20 euros were cost by the ticket

(3.31) a. I owned that car for years

b. That car was owned by me for years

Copulars are particularly difficult to passivise (3.32)-(3.33) – "resembles" outnumbers "is resembled by" by a factor of over 16 thousand in the Google directory. However stative verbs of attitude seem to be an exception (3.34).

(3.32) a. The project proved a bad investment

b. #A bad investment was proved by the project

(3.33) a. The animal resembles a crab

b. ?A crab is resembled by the animal

(3.34) a. The public at large admired her

b. She was admired by the public at large

Beyond this, scope sensitive structures can interfere with the active-passive equivalence, producing sentences that change the semantics of a proposition (Quirk et al., 1985, sec. 3.72). Modal lexemes can produce this effect, where the attitudinal predication moves from one argument to another (*want* from *I* to *you* in 3.35b, unless the proposition is passivised 3.35c), or becomes ambiguous (was Jim or his friend being deliberate in 3.36b?) (see also Weiner and Labov, 1983, pp.29-30).

(3.35) a. I want to marry you

b. @ You want to be married by me

c. ?To be married by you is wanted by me

(3.36) a. Jim deliberately lost his friend at the concert

b. @ His friend was/got deliberately lost by Jim at the concert

Similarly quantifying expressions can cause unacceptable passives (3.37b), necessitating a change in polarity (3.37c), or can change the meaning of a proposition – in (3.38a) most of what Hemingway drinks is beer, while (3.38b) can be interpreted as saying that most beer in existence was consumed by one man.

(3.37) a. He didn't find anything

b. #Anything wasn't found by him

c. Nothing was found by him

(3.38) a. Hemingway mostly drank beer

b. @ Beer was mostly drunk by Hemingway

There are also restrictions on the nature of the arguments a passive takes. As Palmer points out, passive subjects are both animate and conscious more often than not, but not causing – he characterises the passive as typically involving the promotion of a non-Agent and the demotion of an

Agent (Palmer, 1994, pp.117,124). Experiencers are also possible passive *by*-objects according to Biber et al. (1999, p.935), though this can be seen as an alternative formulation of the point made in (3.34), that stative attitudinal verbs passivise readily (since the canonical subject of these verbs is usually an Experiencer). These views are consistent with Jackendoff's proposal of a thematic hierarchy (see page 44).

A principle function of the passive is pragmatic. Using the passive allows canonical objects to occupy the topic (pre-verb) position, and the canonical subject to occupy the post-verb position (3.39). These positions are strongly associated with given and new information, respectively (Biber et al., 1999, p.941). In the case of passives lacking a *by*-phrase, where the canonical subject is unimportant, unknown or understood, the verb phrase occupies the new information position (3.40). Weiner and Labov (1983, p.40-43) also note that these pragmatic patterns are modulated by stylistic factors – the passive is used more in careful speech than casual speech, and more by females than by males.

(3.39) Do you know Tracy's in hospital? $\text{She}_{\text{given}}$ was run over by a bus_{new}.

(3.40) My name is Robert, and I_{given} was diagnosed in $\text{September}_{\text{new}}$.

Passives are also often used to put long clumsy active subjects ("heavy constituents") (3.41a) in the end position (3.41b). Biber et al. (1999, p.940) find a strong tendency for passive objects (canonical subjects) to be longer than their subjects, presumably due to this "end-weighting" principle. Finally, putting the canonical object in the subject position allows it to be elided in coordinated clauses (Palmer, 1994, p.136), as 'he was' is in (3.42).

(3.41) a. $[\text{Those two people we were talking to on the bench in the park yesterday afternoon}]_{\text{subj}}$ found your wallet

b. Your wallet was found by those two people we were talking to on the bench in the park yesterday afternoon

(3.42) He was tarred by his classmates, and then [~~he was~~] feathered by the teachers

These pragmatic considerations lead to strong biases in passive realisation across genres. Overall, passives without a *by*-phrase outnumber those with by a ratio of 4:1 (Quirk et al., 1985, 3.71). Passives of all types are much more frequent in technical and journalistic than in literary texts (Quirk et al., 1985, 3.73, Biber et al., 1999, p.937). An analysis of incidence of passive per sentence in the Susanne sub-corpora ranges from 60% for academic/technical texts, and 25-30% for journalism and *belles lettres*, to under 10% in popular fiction. Biber et al. (1999, p.938) attribute this to the increased complexity, and accordingly length, of noun phrases in formal writing, and the frequent use of impersonal generic statements.

3.2.1.3 The German Passive Construction

The basic word order of German is Subject–Verb–Object. Though case is not always explicitly marked (for example proper nouns and common nouns appearing without an article or adjective do not have distinctive nominative/accusative forms), the strong case marking of noun phrases permits their extensive reordering. Movement is subject to the constraint that one and only one dependent constituent appears before the finite verb, in the so-called *Vorfeld*. In the second clause of (3.43) the dative object *dir* moves into the *Vorfeld*, displacing the subject to a post-verbal position. Similarly

in (3.44) an accusative object *Das Auto* displaces the subject *ich*. Adverbials can also occupy the *Vorfeld*.

(3.43) ... mach dir$_{dat}$ keine Hoffnungen$_{acc}$; dir$_{dat}$ gebe ich$_{nom}$ nichts$_{acc}$!
... make you no hopes; you give I nothing
'... don't raise your hopes, I'll give you nothing!'[26]

(3.44) Das Auto$_{dat}$ habe ich$_{nom}$ 1992 neu gekauft ...
the car have I 1992 new bought
'I bought the car new in 1992 ....'[27]

Since active verb structures allow reordering, often for discourse reasons such as cohesion with preceding text, emphasis of particular sentence elements, or signalling of the newness or givenness of information, the passive structure is not as common in German as it is in English. The German passive is primarily used when the canonical subject is either unknown or unimportant (Durrell, 2002, pp.476-480).

The German *Vorgangspassiv* 'process passive' is very similar in form and in function to the English dynamic passive. As in English, the subject and direct object of a truth-equivalent active sentence (3.45a) move either side of the verb, the canonical subject being embedded in a prepositional phrase (3.45b). The main verb becomes a past participle and follows the auxiliary *werden* 'become'. Again, as in English, the prepositional phrase carrying the canonical subject is optional (3.45c).

(3.45) a. Sie$_{subj}$ hat einen Zimmer$_{dobj}$ hier schon reserviert.
she has a room here already reserved
She has already booked a room here.

b. Ein Zimmer$_{subj}$ hier ist schon von ihr$_{dobj}$ reserviert worden.
a room here is already from her reserved become
A room here has already been booked by her.

c. Ein Zimmer$_{subj}$ hier ist schon reserviert worden.
a room here is already reserved become
A room here has already been booked.

Transitive events lend themselves well to the passive. Stative verbs do not (e.g. *kosten* 'cost', Pollard, 1994, p.292), and psychological verbs (e.g. of perception, belief and experience) vary in their behaviour. Accordingly subjects that are volitional, causative or personal tend to passivise well (Zifonun et al., 1997, p.1796-7). The primary functions of the passive are text structuring (that is to focus and defocus the canonical object and subject respectively; Fleischer et al., 1983, p.197), and to put emphasis on the process of an action (Dreyer and Schmitt, 1985, p.101). The German passive is thus close to Keenan's universal characterisation of the passive (see page 74).

Besides the *Vorgangspassiv,* regarded as the most central and common form (Zifonun et al., 1997, p.1791-3), German also has a much less frequent *Zustandspassiv* 'stative passive' (Durrell, 2002, p.307) which mirrors its English cousin in using the auxiliary *sein* 'be' (3.46), describing resulting states without focusing on the preceding action (Dodd and Zojer, 2003, p.103), and in not having a truth-equivalent active counterpart. A related impersonal passive structure (the 'middle voice') also exists (3.47), which has no clear structural equivalent in English.

[26] German Translation of Fyodor Dostoevsky, 1867: *The Gambler* [Guthenberg Project: http://gutenberg.spiegel.de/dostojew/spieler/toc.htm]

[27] *Das Spiegel* magazine, Second hand car review *Die treusten Weggefährten* [http://www.spiegel.de/auto/aktuell/0,1518,425505,00.html]

(3.46) Ein Zimmer$_{subj}$ hier war schon reserviert.
a room here was already reserved
A room here was already booked.

(3.47) Hier werden Zimmer$_{dobj}$ selten reserviert.
here become rooms seldom reserved
Rooms aren't booked here often.

In this experiment only the process passive is considered, as it forms a true diathesis pair with its truth-equivalent active counterpart.

3.2.1.4 The Chinese *ba* and *bei* constructions

The Chinese *ba* 把 and *bei* 被 coverb constructions have certain similarities to passive constructions: they are alternatives to a less marked active sentence; they involve giving additional prominence to the canonical object; and they are associated with transitivity or affectedness.

Functionally, coverbs are roughly equivalent to prepositions in German and English, but they are full verbs that have a secondary use as argument phrase markers. Their closest equivalent in English is the use of non-finite present participle forms to introduce adverbial phrases – e.g *using* in "He broke into the shop using a crowbar". Generally coverb phrases are placed following the topic and subject positions[28], and immediately before the verb. In the following constructed example four coverb phrases are placed between the subject *Zhangsan* and main verb *touzoule* 'stole'. Coverb phrases, like other arguments, can be topicalised[29] to appear before the subject position and verb.

(3.50) Zhangsan$_{subj}$ **zai**$_{cvrb}$ huochezhan **weile**$_{cvrb}$ zheng qian **gei**$_{cvrb}$ ta erzi shang
Zhangsan **be-located** train-station **act-PFV** earn money **give** he son attend
daxue **yong**$_{cvrb}$ kache **touzoule**$_{verb}$ wo de zixingche$_{dobj}$
university **use** lorry stole-away-PFV me ATTR bicycle
'Zhangsan stole my bike with a lorry at the railway station in order to earn money for his son to go to university.'
张三在火车站为了挣钱给他儿子上大学用卡车偷走了我的自行车

The two constructions differ in that the *ba* construction leaves the subject in its canonical position (3.51a is a basic active Subject–Verb–Object sentence), placing the object before the verb in the *ba* coverb phrase (3.51b).

[28]The category of subject is controversial in Chinese linguistics. The category of topic carries many of the pragmatic functions of the English subject (cf. Keenan, 1976). The subject can be defined as the the argument that in the unmarked condition, and in the absence of an explicit topic, will appear at the head of the clause (in such a case the subject is a putative topic). The Chinese subject is associated with agency (Li and Thompson, 1997, pp.19, 87). Certain locative sentences have no subject. Nor do sentences which take expletive subjects in English (e.g. *it's raining*, *there is/are*). LaPolla (1993, pp.1,4) has argued against the existence of grammatical function in Chinese altogether: "... there have been many attempts to define a subject for Chinese, though no one has succeeded in this venture. ... there has been no grammaticalization of syntactic functions in Chinese. The correct assignment of semantic roles to the constituents of a discourse is done by the listener on the basis of the discourse structure and pragmatics."

[29]Topicalisation in Chinese is a general process, by which any argument or adverbial can be placed at the head of a clause (3.49) – a position usually occupied by the subject (3.48).

(3.48) wo$_{subj}$ chi le san wan rousimian$_{dobj}$
me eat PFV three bowl meat-strip-noodle
'I ate three bowls of pork noodles'
我吃了三碗肉丝面

(3.49) san wan rousimian$_{dobj,topic}$ wo$_{subj}$ chi le
three bowl meat-strip-noodle me eat PFV
'I ate three bowls of pork noodles'
三碗肉丝面我吃了

(3.51) a. tamen$_{subj}$ qieduan le dianli$_{dobj}$
they cut PFV electric-power
'They cut the power'
它们切断了电力

b. tamen$_{subj}$ ba dianli$_{oblq}$ qieduan le
they BA electric-power cut PFV
它们把电力切断了

In the *bei* sentence the subject is displaced into the *bei* coverb phrase and the object moves into the topic position at the head of the sentence (3.53a). The *bei* construction can also drop the canonical subject, in a similar way to the English and German passives, if the Agent-like argument is unknown, unimportant or understood (3.53b). However in such cases the sentence (3.53c) – a common sentence configuration formed by dropping[30] the subject, and topicalising the object – might also be appropriate. Choosing to use the *bei* construction stresses agency; (3.53b) would not be felicitous for a situation in which an electrical fault caused the electricity to cut out (compare the difference between the English *it broke* and *it got broken*).

(3.53) a. dianli$_{dobj,topic}$ bei tamen$_{oblq}$ qieduan le
electric-power BEI they cut PFV
电力被它们切断了

b. dianli$_{dobj,topic}$ bei qieduan le
electric-power BEI cut PFV
电力被切断了

c. dianli$_{dobj,topic}$ qieduan le
electric-power cut PFV
电力切断了

Bei (literally 'cover'[31]) is particularly common in technical texts (McEnery and Xiao, 2005). In written contexts *bei* is equally as common as *ba* (literally 'grasp'), but in spoken contexts approximately only a third as common.[32]

Bei and *ba* share many structural and functional features (Li Dejin and Cheng Meizhen, 1988, pp.494ff; Li and Thompson, 1997, pp.,468ff). Both only apply to transitive verbs that have a resultative complement or aspectual marker of some kind (such as the perfective marker *le* in 3.51 and 3.53; or the resultative complement phrase *wei ri, e, yi, san guo wenzi* in 3.54).[33]

(3.54) ji zhuzuo$_{dobj}$ hui bei yi$_{verb}$ wei ri, e, yi, san guo wenzi
such works can BEI translate make Japan, Russia, Italy, three country text

[30]Chinese allows arguments to be dropped, if it is recoverable from the context (i.e. in similar circumstances to those that would allow pronominalisation in a language like English or German). Similar to the Chomskyan notion of pro-drop, it applies to any argument (not just the subject), and there is no residual marking on the verb or elsewhere (e.g. Spanish and Italian allow pro-drop, but the verb retains marking for number and person). The sentence below, consisting of a verb alone, is perfectly felicitous in an appropriate context (e.g. one where someone has just accused someone else of losing a possession):

(3.52) mei diu
not-PFV lose
'(I) didn't lose (it)'
没丢

[31]*bei* is perhaps unique among coverbs in that its verb interpretation 'cover' is unrelated to its coverb usage, and very rare. In the Academia Sinica Balanced Corpus *bei* appears only five times as a verb, but over 5000 times as a coverb.

[32]The difference in distribution across registers is significant: $p \leq 0.001, \chi^2 = 280$. Data from the Academia Sinica Balanced Corpus, CKIP Group (1998), [http://www.sinica.edu.tw/ftms-bin/kiwi1/mkiwi.sh?language=1], viewed August 2006.

[33] Yip Po-Ching and Rimmington (1997, p.25) suggest that a primary function of the *ba* construction is to leave the post-verb position free for such complements and markers.

‘Such works should be translated into the languages of the three nations, Japan, Russia and Italy’
其著作曾被譯為日、俄、義三國文字。。。[34]

Both express the notion of “disposal” (the original Chinese term is *chuzhi* 处置 meaning ‘handle, deal with, dispose of, take charge of’), that involves affectedness to the canonical subject, whether it describes “something happening to the object” (Li and Thompson, 1997, p.468), or “intentional manipulation or unintentional intervention” towards the object (Yip Po-Ching and Rimmington, 1997, p.93). Both involve the direct object having definite reference (or alternatively specific or generic reference, Wang Hai, 1992, p.526).[35] And both constructions are generally limited to verbs that express some sort of transitive action, so verbs of perception, thinking and movement are usually ruled out. According to some authors the *ba* construction expresses particularly physical or concrete affectedness, though this is by no means absolute – consider Li and Thompson’s example below:

(3.57) ta ba nei jian shiqing liaojie de hen touche
he BA that CL issue understand ATTR very thorough

‘S/he understands that matter very thoroughly’
他把那件事情了解的很透彻

Bei is also carries an adversative connotation (i.e. of negative affect), particularly for the canonical object (3.58a), though this seems to be changing (consider 3.59a).[36]

[34]Academia Sinica Balanced Corpus, CKIP Group (1998), [http://www.sinica.edu.tw/ftms-bin/kiwi1/mkiwi.sh?language=1].

[35]Chinese does not have explicit marking of definiteness (Chen Ping, 2004, pp.1143,1146,1177): “... there is no simple, fully grammaticalized marker of definiteness in Chinese, like the definite article in English. ... Furthermore, it is not obligatory to mark a nominal expression as either definite or indefinite ... [for bare NPs and cardinality expressions] the features of definiteness and indefiniteness cannot be uniquely and unambiguously specified ...”.

Some categories of words are inherently identifiable, such as proper nouns and pronouns. However, for bare and cardinality expressions, which Chen Ping terms “indeterminate expressions”, there is a “strong but seldom absolute” correlation between definiteness and sentence position. Generally, preverbal constituents are definite and postverbal constituents indefinite. Examples (3.55, 3.56) demonstrate this with a minimal pair. Strongly transitive objects are equally likely to be definite or indefinite, and in ditransitive constructions, the first object is more likely to be definite than the second (Chen Ping, 2004, p.1171).

(3.55) keren lai le
guest come PFV

‘The guest(s) has/have arrived’
客人来了

(3.56) lai le keren
come PFV guest

‘A/some guest(s) has/have arrived’
来了客人

[36]Li and Thompson (1997) attribute the change in use to influence from Western languages. They quote one author, Kierman (1969 pp.74-75) who is particularly scathing of the use of *bei* in non-adversative contexts: “There has been a great deal of translation from foreign languages into Chinese during the past half century, including a perfect flood of Marxist material, which the Soviets translated and sold far below cost and which had a profound and continuing impact upon Chinese intelligentsia. The great majority of the translators were hacks, equipped with neither any real linguistic sophistication nor even a very secure grasp of the languages involved and their stylistic niceties.”

(3.58) a. tamen de zinu$_{\text{dobj,topic}}$ ye bei dangdi xuexiao$_{\text{oblq}}$ jujue$_{\text{verb}}$
they ATTR children also BEI local school reject
'Their children were also refused admission to the local school'
他們的子女也被當地學校拒絕[37]

b. dangdi xuexiao$_{\text{oblq}}$ ye jujue$_{\text{verb}}$tamen de zinu$_{\text{dobj,topic}}$
they ATTR children also BEI local school reject
'The local school also refused admission to their children'
當地學校也拒絕他們的子女

(3.59) a. yesheng daxiang$_{\text{dobj,topic}}$ ... bei cunmin$_{\text{oblq}}$ yong watuji jiuchu$_{\text{verb}}$
wild elephant … BEI villager use digger rescue
'Wild elephant rescued by villagers with digger'
野生大象。。。被村民用挖土机救出[38]

b. cunmin$_{\text{oblq}}$ yong watuji jiuchu$_{\text{verb}}$ yesheng daxiang$_{\text{dobj}}$
villager use digger rescue wild elephant
'Villagers rescue wild elephant with digger'
村民用挖土机救出野生大象

There is still little consensus among formal syntactic accounts of these constructions. Tsao Feng-fu (1987) sees *ba* as a pragmatic marker of "secondary topic", while Bender (2000) sees it as a fully fledged verb, whose own valency properties account for its distribution. For Liu Feng-hsi (1997) *ba* is a marker of bounded aspect, which determines the situation type of the verb to which it attaches through by isomorphism with its specific (and so bounded) object, in a manner similar to Verkuyl (1993). For Ting Jen (1998) *bei* sentences represent three different constructions, each with a distinct structure and derivation.

3.2.1.5 Coding of Semantic Features

As has just been seen, a range of syntactic, semantic and discourse/pragmatic features are thought to operate in licensing the diatheses of interest. The materials used in these acceptability experiments were hand coded with such features, both of the predicate itself (and by extension the event denoted) and of each individual argument (and so the nature of its referent's involvment in the event).

The semantic feature on which the prior hypotheses are based is animacy. This was chosen for two reasons. First of all, animacy is a well established ontological category, which has external support in theories of biology (as will be seen below, animacy is interpreted in this strict, biological sense). Secondly, animacy (or features for which it may be a prerequisite, such as agency, benefit, possession, awareness) is claimed to be a determining factor in the productivity and acceptability of all the diatheses described here.

A series of further features suggested by the literature were also coded by hand. Those that later proved to be of interest (by demonstrating a significant effect on diathesis acceptability) are elaborated in the subsections that follow. Features that were coded but that did not reveal interesting and significant patterns include: boundedness of the argument referents; length of the argument phrase in words; an argument being a product or resource in the event (Fillmore's Factitive role);

[37] Academia Sinica Balanced Corpus, CKIP Group (1998), [http://www.sinica.edu.tw/ftms-bin/kiwi1/mkiwi.sh?language=1].

[38] Web news headline, WhNews.cn [http://whnews.cn/2006-07/15/content_939957.htm].

receipt by one participant of another participant; and physical contact between participants. These features are not discussed further here.

While some of the coded features might be considered to be graded properties (e.g. someone can be more or less willingly involved in a situation) they were generally given only binary values. Affect was coded with one of three values: positive, neutral and negative. The large number of individual coding judgements made (over 4000), and the requirement for expert coders, made a full verification of the coding impractical, but small informal verifications were carried out, as described below for animacy.

As discussed in the introduction (see page 8), semantics is being treated at the level of the event proposition that underlies an utterance, and without regard for its ultimate truth, or the degree of confidence expressed in it by the speaker.

Animacy In all the acceptability experiments described in this section, test sentences were selected so that equal numbers of each would have an animate and inanimate direct object (in the case of passive and *ba/bei* constructions) or prepositional object (in the case of the English dative/benefactive). Animacy of the subject would also be an interesting feature to control for, but adequate numbers of such sentences were not found among the authentic film script materials used here.

The definition of animacy is not entirely uncontroversial – it may only include conscious intelligent life (humans, and perhaps a number of higher animals), or be extended to all life (for example grass, and amoebae). In particular, humans, social groupings, organisations, animals and machines are often viewed as being animate to varying degrees (as discussed earlier in sections 2.3 and 2.4). The fact that organisations and other groups are often viewed as representing their constituent members, means that English singular nouns that represent groups of people can sometimes be given plural (3.60b), or mixed agreement on the verb (3.60c) (this mismatch between "notional" and "grammatical" number is common in British English, see e.g. Levin, 2006; in American English it is rare, Bock et al., 2004, 2006).

(3.60) a. Manchester United$_{sing}$ is$_{sing}$ without a doubt the most popular team for the most popular sport.[39]

b. Manchester United$_{sing}$ are$_{plur}$ a team in transition struggling to rediscover past glories.[40]

c. Manchester United$_{sing}$ are$_{plur}$ a club that strongly believes$_{sing}$ in the development of youth.[41]

In a small informal elicitation exercise, colleagues were presented with eight sentences, each of which implied the sentience or volition of the subject (e.g. *The cockroach noticed it had been spotted*, or *The region expelled some of its people*). The subject noun phrases were taken from English experimental materials that were judged to be borderline cases for animacy coding. All the sentences were in the simple past tense, so plural/singular agreement did not play a role. Some noun phrases were altered to make them more generally interpretable: e.g. *region* replaced *Middle Earth* (a region in Tolkien's fantasy world), and *ship* replaced *Millennium Falcon* (a space ship in the *Star Wars* series). Five colleagues responded (three Irish English speakers, one U.S English speaker, and one Dutch/Irish English bilingual), rating each constructed sentence for semantic acceptability on a

[39] [http://www.faqfarm.com/Q/What_is_the_most_popular_sports_team_in_the_world], viewed August 2006.
[40] [http://worldsoccer.about.com/od/fac2/a/facupmangun05.htm], viewed August 2006.
[41] [http://www.manutdsoccerschools.com/index.html?hongkong.htm], viewed August 2006.

five point Likert scale (from +2 for "perfect" to -2 for "terrible"). The results were mixed. Some subject noun phrases received uniformly positive (*the British*) or negative ratings (*my house*). All the other phrases (*cockroach, team, main office, work, region, ship*) elicited a mix of positive and negative ratings, and this lack of unanimity was present even among the three monolingual native Irish English speakers.

As a result of these differences of opinion, a rather strict and literal approach was adopted when coding argument phrases as animate: only accepting nouns that exclusively refer to animate individuals or groups of animate individuals (among the examples here, *British* and *cockroach*). Nouns that can either refer to groups or organisations (such as *team*), and nouns that only referred to groups by sense extension (*ship, main office*) were coded as inanimate.

Telicity and Duration Situational aspect (i.e. *Aktionsart*) was considered as an coding feature. The stative/dynamic distinction is thought to play a role in the generation of the passive, and the *bei* and *ba* constructions both have a structural requirement of a resultative complement (which suggests telicity).

However, reliable coding of sentences proved difficult. Aspect typologies differ. Vendler's four-way typology (1967) is perhaps the best established: States (*love, be, resemble*), Activities (*run, snore, fall*), Achievements (*smash, realise*) and Accomplishments (*build, make, destroy*). Both Moens and Steedman (1988) and Smith (1991) posit a fifth aspectual category for short-duration atelic events, like *sneeze.* However they differ in the mode of definition: Smith (1991) uses the binary features *dynamic, durative* and *telic* to characterise aspectual types; while Moens and Steedman (1988) operationalise the categories with linguistic tests (they apply various tense forms and temporal adjuncts to the sentence whose aspectual class is in question). The picture is complicated by aspectual coercion (Verkuyl, 1993). For example a verb like *climb* may be thought to constitute an Activity by default (atelic, durative), but adding an object that refers to an explicit end point (e.g. *climb a mountain*) yields an Accomplishment (telic, durative).

By definition, some aspectual types are less likely. Point (semelfactive) events proved difficult to find among the materials (perhaps because they are rather inconsequential – short and without after-effects). Dative situations involve the (often durative) transfer of Goods or Information to a Goal of some sort, which usually imposes a telic interpretation. Benefactive situations usually involve the completion of some act on the behalf of, or for the good of the beneficiary, and so are also typically telic (3.61 is arguably a rare atelic example).

(3.61) **Charles Foster Kane**: **They've been making statues** for some two thousand years, and I've only been collecting for five.[42]

For these reasons, it was difficult to assemble large enough sample sizes of each aspectual category to yield significant results, and items were annotated instead for the binary features of Durative and Telic (again, Stative examples were too rare to yield detectable effects in the experiments described below).

These features only applied to the experimental clause as a whole. Predicate tokens were considered telic if they had a clear end-point with a relevant change of state. Two definitions were used – a narrow one that only included material states, and a broad one that also included abstract states. There were many borderline cases. For example in the first sentence below, a memory wiping device

[42] *Citizen Kane*, Orson Welles, 1941.

(that flashes) is being discussed. While the flash event is delimited with a clear end point, there is little lasting effect on the device itself.

(3.62) you flash that thing[43]

This second sentence describes an ongoing process (as such not telic), but as a population is being discussed, it describes many telic sub-events, as each person leaves.[44]

(3.63) they are leaving Middle-Earth[45]

This last sentence describes a *forbidding* event, that if it has an effect (that is if it is complied with) would mean that something *doesn't* happen. This was coded as materially atelic, but psychologically telic.

(3.64) I forbid it[46]

Similar difficulties of decision obtained for duration – as all events have an end, they also have duration. Sentences were annotated as having duration if they involved relevant intermediate states or stages. For example *flash* and *salute* are both lack relevant internal structure – one is unlikely to speak of someone being halfway through a flashing or saluting event, in contrast to a *cooking* event, such as (3.65).

(3.65) My mother's gonna make some fried peppers and sausage for us[47]

A different problem is posed by verbs that directly describe a point of transition, which is conventionally preceded or followed by a longer event. For example *sending* usually entails a longer *delivering* event, and *finding* is usually preceded by a search. These were interpreted conservatively, as not having duration.

Affect Negative affect is said to be (or have been) a determining factor in the admissibility of the Chinese *bei* construction. There is also the suggestion that the English *got* passive is associated with negative outcomes for the canonical object. In the situation described below there is a suggestion that *he* comes out the worse for the commercial transaction, or at least in not a completely willing participant.

(3.66) The latter part of last year, **he got sold a computer**. I would like to say, bought a computer, but since he got so much more than he needed, I would have to say that it was an act of (questionably ethical) salesmanship rather than consumerism.[48]

Each argument was annotated for affect on the basis of whether the situation described has or is a positive, neutral or negative outcome for that argument. The sentence as a whole was also annotated for affect, on the basis of it being a good or bad outcome in general.

Affect can be influenced by viewpoint. In the sentence below from the film *Jaws*, one can see being caught as a negative outcome for the man-eating shark, but presumably sympathise with its objective, and so annotate the *catch* situation as positive. In the second example from *Finding Nemo* the eponymous captive aquatic life is the protagonist, so this is viewed as a negative outcome.

[43] *Men in Black*, Barry Sonnenfeld, 1997.
[44] An alternative interpretation is also possible in which *they* have all started the process of leaving, but none has yet left.
[45] *The Fellowship of the Ring*, Peter Jackson, 2001.
[46] *Titanic,* James Cameron, 1997.
[47] *Goodfellas,* Martin Scorsese, 1990.
[48] Weblog, *The Bad Samaritan*, August 2001, http://www.badsamaritan.com/archives/2001/08/.

(3.67) I'll$_{pos}$ catch$_{pos}$ this bird$_{neg}$ for you[49]

(3.68) he$_{neg}$ was taken$_{neg}$ by these divers$_{pos}$[50]

As just discussed, semantics is being treated at the level of the proposition that underlies an utterance, independently of its truth value. Hence the next sentence is annotated as having positive affect, since having a survival suit would have been positive, if it had happened.

(3.69) they$_{neu}$ couldn't design$_{pos}$ a survival suit$_{neu}$ for us$_{pos}$[51]

New/Given and Definiteness The new/given distinction has had various characterisations (see Prince, 1981). The distinction used here is based on what Prince (1981, p.230-231) calls the "Shared Knowledge" approach, in which the speaker is assumed to structure a sentence based on what s/he believes the hearer to know or be acquainted with. Given referents include only those that have been mentioned in preceding text, or that are otherwise perceptually present in the context (e.g. by sight or smell) ("evoked" in Prince's typology). New referents are those that are introduced for the first time in the clause in question, and this includes referents that can be logically related to given referent (what Prince calls *"*inferables*"*, such as *the driver* of a given *bus*; *ibid* p.237).

Procedurally, pronouns and proper names arguments were treated as given, since they signal that the speaker believes the listener to be familiar with them. Common nouns were annotated as given if they were found in the preceding context. Definite articles, demonstrative and personal pronouns and proper names were all treated as definite.

Concrete/Abstract The annotation of arguments was relatively straightforward, following the same ontological approach as that used for animacy. But verbs sometimes presented problems. For example the sentence below involves material action (picking up the receiver, dialling number, as well as the motor actions involved in speaking), but it was treated as an abstract verb, since its primary focus and purpose is the abstract communicative act of transferring information.

(3.70) I rang a few people[52]

Awareness and Volition Arguments that are aware of or volitionally involved in the situation or event represented in the sentence are most often animate. Inanimate arguments can also be aware, if they can be taken to refer to a person or people by holonymic extension, for example from the names of nations to their people:

(3.71) America owes you a debt of gratitude[53]

Verbs of desire are problematic, since a participants volition may be involuntary (consider addicts). Such cases were coded as volitional:

(3.72) I want 50% of ye plunder[54]

[49] *Jaws*, Stephen Spielberg, 1975.
[50] *Finding Nemo*, Andrew Stanton, 2003.
[51] *The Towering Inferno*, Irwen Allen, 1974.
[52] *Four Weddings and a Funeral*, Mike Newell, 1994.
[53] *Forrest Gump,* Robert Zemeckis, 1994.
[54] *Pirates of the Caribbean*, Gore Verbinski, 2003.

Causing and Affected Agency is central to many accounts of linking, but it has differing characterisations. Though volition often patterns with causation, but they are not mutually necessary. In the example above (3.71) the people of *America* presumably agree that they owe Forrest Gump their gratitude (i.e. the people are volitionally involved), but it was rather his actions that are the cause of that. In an example like that below, *sorrow* would be a cause on a literal interpretation, but is not on the interpretation of *I got sad*.

(3.73) I was surrounded by an endless sorrow[55]

In *like* and *hate* type situations, it is often difficult to say which, if either of the participants is the cause. For example is it an attribute of Colonel Saito's personality that makes him dislike his British prisoners, or can they be seen as the cause. In this case *the British* was annotated as the cause, as it is their behaviour (rebelling against their jailers) that is the source of the Colonel's feelings.

(3.74) I hate the British[56]

Non-agents can also be causers, for example by issuing an order to others – in the following example both *you* and *me* are causers, though only *me* is volitional.

(3.75) You will bring Captain Solo and the Wookie to me[57]

Affectedness was given both a narrow (solely material) and broad (material or psychological) interpretation. Destruction and creation are maximal cases of material affectedness:

(3.76) God wrought good men[58]

(3.77) Dinosaurs eat man[59]

But it was often very hard to decide on borderline cases. A *kissing* event does involve (material) movement but this is not its purpose. In a social greeting situation, the psychological effect it causes can be thought of as minimal. However in the example below, the exclamation *yeech*, makes it clear that there was a (negative) psychological effect on *me*.

(3.78) He kissed me, Yeech.[60][61]

In a situation like that below, both *you* and *my father* can be viewed as psychologically affected, the latter as a beneficiary, and the former because they are beholden to him.

(3.79) you served my father for many years[62]

In this case both *father* and *you* were coded as being psychologically affected, while only *you* was physically affected (by having to carry out tasks in that service).

[55] *Crouching Tiger, Hidden Dragon*, Ang Lee, 2000.
[56] *The Bridge on the River Kwai,* David Lean, 1957.
[57] *Return of the Jedi,* George Lucas, 1983.
[58] *Doctor Zhivago,* David Lean, 1965.
[59] *Jurassic Park*, Steven Spielberg, 1993.
[60] *Annie Hall,* Woodie Allen, 1977.
[61] The hash symbol '#' is used to mark sentences that are semantically illformed. See the appendix (page 183) for a list of marks and other notations used. The use of such informal notation of acceptability is discussed in section 1.3.
[62] *Star Wars,* George Lucas, 1977.

3.2.2 The English Dative and Benefactive Diatheses

3.2.2.1 Background

The first acceptability experiment to be described is that on the dative and benefactive diatheses. Animacy, and several semantic categories that may entail animacy (receipt, possession, benefit) have been suggested as attributes of the indirect object, that are necessary for the use of the dative or benefactive double object constructions (see pages 2 and 50). In this experiment, the hypothesis to be tested is whether the animacy of the prepositional object in sentences such as (3.80a) or (3.81a) is a determining factor in the success of the diathesis (3.80b, 3.81b).

(3.80) a. **Axel Foley**: ... get [your supervisor]$_{\text{dobj}}$ [for me]$_{\text{oblq}}$.[63]

b. **Axel Foley**: ... get [me]$_{\text{iobj}}$ [your supervisor]$_{\text{dobj}}$.

(3.81) a. **Scott Evil**: Well my friend Sweet Jay took [me]$_{\text{dobj}}$ [to that video arcade in town]$_{\text{oblq}}$ ...[64]

b. #**Scott Evil**: Well my friend Sweet Jay took [that video arcade in town]$_{\text{iobj}}$ [me]$_{\text{dobj}}$ ...

Authentic sentences which have the configuration Verb–NounPhrase–PrepositionalPhrase, a verb that is known to support the diathesis (according to Levin, 1993, sec. 2.1), and the appropriate preposition (*to* for dative sentences, and *for* for benefactive sentences) will be compared to their manually constructed diathesis variant. The difference between the acceptability scores of the original and altered variants will give a measure of diathesis success for each sentence (termed the "difference score"), and group means of these values will be compared to determine if animacy of the prepositional object has a significant effect.

A post-hoc investigation of additional features (those described in section 3.2.1.5) will also be performed on a similar basis. Semantic (animacy, benefit, receipt) and discourse properties (givenness, heaviness) of the indirect object have been suggested as possible determinants (see section 3.2.1.1). Such features will be compared by the strength of their effect on diathesis success.

The results may also confirm some of the arguments against very early approaches to linking (presented in chapter 1), such as the contention that diathesis forms are not solely licenced by the lexical entry of the verb, or by selectional restrictions.

3.2.2.2 Materials

As discussed in section 3.1.1, textual context has a considerable effect on acceptability judgements, as isolated test sentences can be interpreted in many ways (see section 3.2.6 for a demonstration). Further, utterances used as the basis of linguistic argumentation can also often seem contrived and unnatural. By using authentic materials, both of these issues can be addressed.

If speech is considered a more basic mode of language use than writing, everyday conversations could be expected to make appropriate material. However, spoken dialogue is full of pauses, errors and repairs (3.82), even in rather more formal contexts like media interviews (3.83) and university lectures (3.84).

(3.82) Yeah I think you know he's such a er er he does, he he's very clean but he's not a good housewife.[65]

[63] *Beverly Hills Cop*, Martin Brest, 1984.
[64] *Austin Powers*, Jay Roach, 1997.
[65] BNC text KBF: 62 year old female, English Home Counties.

(3.83) You know, racing has the most fatal death per percentage.[66]

(3.84) What I ... what I thought was, well let's have erm, Woodrow Wilson, okay, as you said at the beginning of the book, Freud admits that he didn't like Wilson, and that he felt betrayed by Wilson[67]

For this series of experiments (sections 3.2.2 to 3.2.5), film scripts were chosen as a source of material, to offer compact everyday examples of language, whose context may be familiar to respondents. Assuming that it is unlikely that participants would remember detailed film dialogues word-for-word, such materials can be subjected to experimental manipulations, while retaining some of their authentic character. Scripts are idealised dialogues, but do constitute authentic language production by the script writer. Thus they offer a compromise between data quality and data authenticity.

To provide familiar contexts, popular films were chosen where possible. For the English experiment, candidate films were primarily selected from rankings of films, with the addition of a number suggested by colleagues. The UK broadcaster Channel 4's *100 Greatest Films*[68] was compiled through online voting, while the *Movie Times* ranking was based on total box-office receipts in the US, adjusted for inflation[69] (as a result *Gone with the Wind* ranks first). From these films a subset of 102 that were familiar to the author were selected.[70] The genres were various, covering cartoons, musicals, science-fiction, comedy, thrillers, drama and romance.

Websites about cinema[71] proved a rich resource for extracts from iconic scenes, as contributed by web users. The extracts are short, semantically coherent and self-contained. They may or may not be identical to the original script, but they do constitute authentic language usage on the part of the web users who submit them to the site. For example one web user submitted variant (3.85b), which was very different in meaning from the original text (3.85a), but nevertheless remained structurally and contextually wellformed.

(3.85) a. **Old Soldier**: The doctor's a gentleman.[72]
Petya: Right! It's written all over him.
Old Soldier: He's a good man.
Petya: God rot good men.

b. ...
Petya: God wrought good men.

Candidate sentences were chosen by hand, following fixed criteria for each of the item categories (fillers, active/passive test items, etc). The materials had been randomised in order, and candidates were chosen in this order, to eliminate experimenter bias. Selection of the dative items was aided by a listing of verbs from Levin (1993) that license this diathesis. For dative/benefactive tests items, finite declarative active sentences with an one overt noun phrase direct object followed by one *for* or *to* prepositional phrase was chosen (3.80). Experimental items were balanced for the animacy of the prepositional object, as described in section 3.2.1.5. Obviously not all the items selected denote

[66] Horse jockey being interviewed on Pat Kenny talk show, RTÉ television, November 2004.
[67] BNC text HE2: Transcript of lecture at Kings College London.
[68] [http://www.channel4.com/film/newsfeatures/microsites/G/greatest/results/control.jsp], viewed August 2006.
[69] [http://www.the-movie-times.com/thrsdir/alltime.mv?adjusted+ByAG], viewed August 2006.
[70] The collected extracted quotes from these 102 films amounted to over 140 thousand words in about eleven thousand lines of text. The amount of material gathered by film varied widely, from almost four and a half thousand words for *Fellowship of the Ring* to just 100 words for *Jungle Book*.
[71] Primarily the *Internet Movie Database* (*IMDB*), [http://www.imdb.com/], viewed August 2006.
[72] *Doctor Zhivago*, David Lean, 1965

a dative or benefactive situation – it is the purpose of the experiment to delimit these classes by the success of the diathesis (e.g. the unacceptability of 3.81b suggests that 3.81a does not denote a dative event).

Some general categories of sentences were excluded from consideration (as both dative/benefactive and passive test materials were collected together, the exclusion criteria for both are described here). Any sentence including language that might be seen as offensive by participants was not used (3.86), as was archaic or other language (so-called *gadzookery*) that participants could not be expected to judge for acceptability – for example (3.87) does not conform to the contemporary English pattern of Subject–Verb–Complement.

(3.86) **Jules:** Marcellus Wallace don't like to be fucked by anybody, except Mrs. Wallace.[73]

(3.87) **Boromir:** By the blood of our people$_{\text{comp}}$ are your lands$_{\text{subj}}$ kept safe.[74]

A series of other structural criteria were used to exclude sentences, because of confounding interactions they may have with diatheses. Negative quantifiers and ellipsis (3.88) can interact with sentence position, so such sentences were excluded (see page 77).

(3.88) a. **Rhett Butler**: None of the fools$_{\text{subj}}$ you've ever known have kissed you$_{\text{dobj}}$ like this, have they?[75]

b. ***Rhett Butler**: You$_{\text{subj}}$ have ever been kissed like this by none of the fools$_{\text{oblq}}$ you've ever known, have they?

Modal verbs (other than tense auxiliaries) and attitudinal adverbs (3.89) and anaphora between arguments (3.90) are also affected by sentence position.

(3.89) a. **Marcus Rathbone**: What's more, the captain$_{\text{subj}}$ deliberately struck me$_{\text{dobj}}$ across the face![76]

b. @**Marcus Rathbone**: What's more, I$_{\text{subj}}$ was deliberately struck across the face by the captain$_{\text{oblq}}$!

(3.90) a. **Dr. Peter Venkman**: Somebody$_{\text{i,subj}}$ blows their$_{\text{i}}$ nose$_{\text{dobj}}$ and you want to keep it?[77]

b. **Dr. Peter Venkman**: Their$_{\text{i}}$ nose$_{\text{subj}}$ is blown by somebody$_{\text{i,oblq}}$ and you want to keep it?

Idiomatic (3.91) and phrasal verb sentences were also judged to be less amenable to diatheses, as were rhymes (3.92).

(3.91) a. **Black Knight**: Right, I'll do you$_{\text{dobj}}$ for that$_{\text{oblq}}$![78]

b. ***Black Knight**: Right, I'll do that$_{\text{iobj}}$ you$_{\text{dobj}}$!

[73] *Pulp Fiction*, Quentin Tarantino, 1994.
[74] *Lord of the Rings*, Peter Jackson, 2003.
[75] *Gone with the Wind*, Victor Fleming, 1939.
[76] *Airport*, George Seaton, 1970.
[77] *Ghostbusters*, Ivan Reitman, 1984.
[78] *Monty Python and the Holy Grail*, Terry Gilliam, Terry Jones, 1975.

(3.92) a. **Willy Wonka**: [singing] In springtime, the only pretty ring time, $\text{birds}_{\text{subj}}$ sing hey ding, a-ding, a-$\text{ding}_{\text{dobj}}$.

b. **Willy Wonka**: [singing] In springtime, the only pretty ring time, hey ding, a-ding, a-$\text{ding}_{\text{subj}}$ is sung by $\text{birds}_{\text{oblq}}$.

General fillers were chosen arbitrarily. A small number of related fillers were chosen in the marked form of the diathesis – that is double object dative and benefactive sentences.

For the rating scale used (magnitude estimation) a reference item must be chosen. To ensure that some sentences would get a higher rating than the reference, an original and wellformed, but slightly unwieldy sentence was chosen for the reference (3.93).

(3.93) **Alvy Singer**: I can't go into a movie that's already started[79]

Five practice sentences were also arbitrarily chosen, three of them wellformed (3.94-3.96) (the last is wellformed in its context, in which Kev has asked for a pizza). The other two were manually altered to yield an illformed sentence (3.97-3.98). Three of the practice sentences included structures of interest in the experiment: benefactive (3.95), passive (3.96) and dative (3.97).

(3.94) **Joe**: ... it may take a little longer [80]

(3.95) **Yu Shu Lien**: I was surrounded by an endless sorrow[81]

(3.96) **Kev**: Did anyone order me a plain cheese [82]

(3.97) **Indiana**: Throw the idol me[83]

(3.98) **Mandy**: Stop thinking sex about[84]

The final set of English sentences had 30 test sentences for the passive diathesis (to be described in section 3.2.3.2), 24 test sentences for the dative/benefactive diatheses, and 33 fillers (14 general, 13 related, 5 practice and 1 reference). As shown in table 3.1 a total 87 test dialogues were assembled. Including both original and altered (i.e. diathesis variant) versions, there were 154 sentence stimuli. The test sentences were evenly split by animacy – 12 of the original dative/benefactives had a prepositional object that was animate, and 12 inanimate.[85]

All sentences were supplemented with surrounding dialogue, and descriptions of film and scene. Diatheses variants were produced for test sentences. An example of such a supporting context can be seen in the instructions, reproduced on page 185 of the appendices.

3.2.2.3 Procedure

To allow large numbers of participants to be recruited, and to automate the storage of responses, it was decided to deliver the experiment via a series of web pages. A custom application was written to control access to the experiment (using a single password shared by all respondents), to record

[79] *Annie Hall*, Woodie Allen, 1977.
[80] *Some Like it Hot*, Billy Wilder, 1959.
[81] *Crouching Tiger,* Hidden Dragon, Ang Lee, 2000
[82] *Home Alone*, Chris Columbus, 1990.
[83] *Raiders of the Lost Ark*, Steven Spielberg, 1981.
[84] *Life of Brian*, Monty Python, 1979.
[85] As mentioned in section 3.2.1.5, a classification by aspectual types was attempted, but proved unsatisfactory. However, the provisional classification by aspectual types was: 3 Process items, 7 Culminated Processes and 14 Culminations, using the terminology of Moens and Steedman (1988).

Table 3.1: Summary of experimental materials for English acceptability experiment

	Original	Altered	Total
General Fillers	14	-	14
Related Fillers	13	13	26
Reference/Practise	4	2	6
Passive Test	30	30	60
Dative/Benefactive Test	24	24	48

information about each respondent as entered into a pre-experiment survey, and to present a random but balanced subset of the materials to each, recording results as plain text files on a server.[86]

For the English experiment a paper based pilot was carried out first. This was followed by a web-based pilot with 17 participants, to validate the methodology, software, materials and instructions. Some minor errors in materials were discovered during piloting, and the instructions underwent several revisions in response to pilot feedback. One major revision was that calibration exercises with other stimulus modalities were removed. Bard et al. (1996) recommend that users be given a magnitude estimation task on some physical continuum, to introduce the concept before applying it to the non-physical continuum of language acceptability. Several variations on the instructions were tried out during the pilot, including line length and brightness judgement tasks, but pilot respondents reported that these made the instructions longer, less coherent, and did not aid in their understanding of the method.

For the live experiment participants were initially recruited among friends, colleagues and acquaintances via email. The email asked people to pass on the message to other acquaintances. The use of a password ensured respondents were limited to those that had been passed an email, but could not in itself prevent repeated or facetious user sessions. A second tranche of respondents was recruited via postings to online forums that had cinema as their theme (this second stage of the experiment was performed in parallel with two variants that used a different judgement scale, but for respondents assigned the magnitude estimation scale, the experiment was identical to that presented to the first group of respondents – see section 3.2.7 for details of the other judgement scale experiments). Checks were made at the analysis stage to see that users were making a substantial responses in line with the instructions (described in the results section below).

Before respondents took part in the experiment, they were asked for some non-identifying personal data: gender, handedness (personal and immediate relatives) and age band. Participants were also asked if they played a musical instrument, how often they watched a film, and if they had received training in any of four professional categories, that might be expected to result in "expert" opinions on language: linguist; school or language teacher; psychologist; or author/editor ("expert" participants may give systematically different responses than non-experts – see section 3.1.2).[87] Finally, participants were asked to identify what dialect of the language they spoke, and if it was their native tongue. For non-native speakers, the number of years spent learning English, number of years living or working in an English-speaking environment and percentage of the time they spoke English now were asked for, as well as their native language.

The judgement scale used was magnitude estimation (see section 3.1.2.3). To introduce the

[86]The application was programmed in Perl, using the CGI interface and running with the Apache web server software on a Sun Solaris platform.

[87]Unfortunately, general level of education was not asked for. This would have proved useful in investigating the effect of numeracy on the results received.

scale, four toy sentences were given, together with a suggested marking strategy. The first, slightly circuitous sentence, (3.99a) was presented as the reference sentence with a reference score of 100, followed by a "perfect" sentence (3.99b) (suggested judgement: twice as good as reference; score 200). The last two sentences exemplified a clear syntactic error (3.99c) (twice as bad as reference; score 50) and a scarcely interpretable one (3.99d) (one third as good as reference; score 30). The full instructions are reproduced in the appendices, on page 185.

(3.99) a. The mat was what the cat sat on

b. The cat sat on the mat

c. The cat sitted on the mat

d. The cat sat on mat the

Participants were told that they could choose any reference score that they liked, and that other scores could be any whole number, fraction or decimal, but that zero or negative numbers could not be used. Participants were asked to imagine that they were helping a friend who was learning English and that the criteria for judging a sentence better or worse in the English experiment were whether it was "English-sounding ... like a native speaking (ignoring punctuation)" or was "strange and unnatural [sounding]". The were explicitly told not to consider punctuation or any prescriptive grammars, and that spoken patterns such as slang, hesitations and repairs were acceptable as long as they sounded authentic. It was stressed that participants should not spend much time on the decision and should instead rely on their "first impressions" and "gut-reaction".

Participants were told the test item would appear in its surrounding dialogue, and with background descriptions. The test item itself would be highlighted with an underline. Participants were to read the scene and film descriptions and indicate if they were already familiar with the film and scene in question. At the end of the instructions a worked example was given, together with a synopsis of the instructions:

- Give the reference quote any number you want (it will be repeated at the start of each page as a reminder).
- Score all the other quotes relative to this one, giving big numbers for English-sounding ones and small numbers for strange or unnatural ones.
- Try to make up your mind as quickly as possible without rushing. There are no 'right' or 'wrong' answers – it is your first impression we are interested in.
- Use as wide a range of numbers as you want.
- Remember to mark 'yes' or 'no' to tell us if you are familiar with the scene and film.
- Base your score only on the underlined portion of the dialogue.
- Please go straight through the questionnaire, without using the 'back' or 'reload' buttons on your browser.

Once a participant started the experiment, s/he was presented with three script excerpts to a screen.[88] Participants were not allowed to progress to the next page without having scored all three items and

[88] For the second recruitment group from web forums, excerpts were presented one to a page.

having answered the multiple choice questions on being familiar with the relevant films and scenes (in the follow on experiment on measurement scales, only one item was presented per page). The first item presented was the reference item, and this was to be given an arbitrary score. Every following item was to be scored relative to this, and the reference sentence and score were repeated on each page as a reminder, together with the bullet point instructions above.[89] The first five items seen after the reference were practise items, though they were presented in an identical manner to the experimental items that followed them.

Participants were asked to judge 30 items including the reference and practise items.[90] This subset of items was selected randomly from the materials, and ordered randomly. This randomisation was constrained to ensure that the proportion of each category of item (e.g. general fillers, related fillers, active sentences with an inanimate object, etc) presented to each participant followed the proportions in the materials as a whole, and that the categories were distributed evenly across the sequence of items. Participants could take breaks at any time, and could also halt the experiment at any point by closing their browser. Participants were not prevented from going back (this cannot be prevented in web pages), but if scores were changed, this was recorded in the server records, and such sessions were excluded from analyses of the data. Participants who completed a set of 30 items were thanked and given the option of continuing on, to view another 30. No participant saw the same item twice, or two variants of the same item (e.g the active and passive versions of the same sentence).

3.2.2.4 Results

Not all 161 sessions recorded were considered valid for analysis. Non-native respondents were excluded (9 sessions).[91] A session was also deemed invalid and excluded from subsequent analysis if participants did not carry out the task as described, by dropping out within the initial six warm-up sentences (62 sessions), or by backtracking to change responses (2 sessions). If respondents gave a lower average response to the two intentionally unacceptable practise sentences (3.97-3.98), than they did to the three attested and acceptable sentences (3.94-3.96), they were excluded for having inverted the judgement scale (one session).

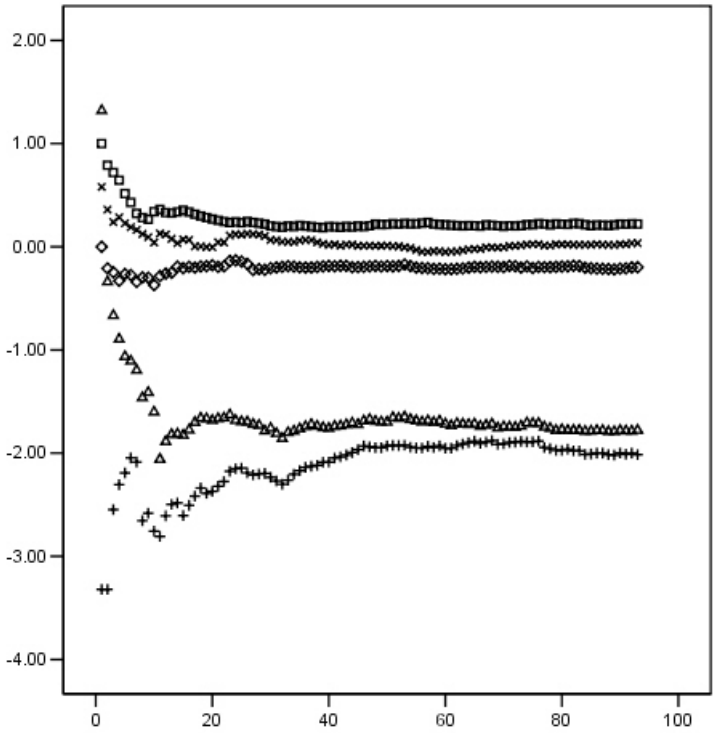
Extreme outlier judgements were also examined to see if they were facetious responses, and two suspect sessions were identified. The first of these respondents gave the reference sentence a score of one, and the authentic practise sentences scores of 5, 10 and 25. The altered dative sentence *you don't even send my house a dress* (which received a marginal score overall) was given a score of 11954. The second suspect session started by giving apparently reasonable responses, but then started entering successively larger scores such as 4869876, 484879279827 and 204534034504530433 (this last score was for the sentence *We've analyzed their attack*, relative to a reference score of 100). These

[89] Using the magnitude estimation technique, the proportional scores given are transitive, so results should be identical if participants make their judgements relative to the previous item, rather than the reference item. However, such scores may drift over the course of the experiment. As a precaution, participants were asked to make all judgements relative to the reference.

[90] For the second recruitment group from web forums, participants were asked to judge 15 items.

[91] Non-natives were originally included in the study so that the effect of level of second language competence, and of interference with the first language could be examined. However the numbers of non-natives were not large enough to allow these analyses. Their responses were excluded from the substantive analysis because they were expected to introduce excessive variation in the data. Non-natives gave higher scores, with a smaller data spread, meaning that they were less opinionated, or less confident in their opinions on the acceptability of sentences (the effect for altered sentences was significant; two-tailed t-test of independent samples over individual responses, $p < 0.05$). Surprisingly linguists did not mark significantly differently from non-linguists, and the number of other language "experts" (teachers, authors, editors) were too small to draw any conclusions.

Figure 3.1: Convergence of mean judgement for practice sentences, as responses were received



The X axis represents the number of responses, the Y axis the mean score up to that point.
Symbols represent the following sentence stimuli (see page 92):

- squares (3.94): ... *it may take a little longer*
- diagonal crosses (3.96): *I was surrounded by an endless sorrow*
- diamonds (3.95): *Did anyone order me a plain cheese*
- triangles (3.97): *Throw the idol me*
- crosses (3.98): *Stop thinking sex about*

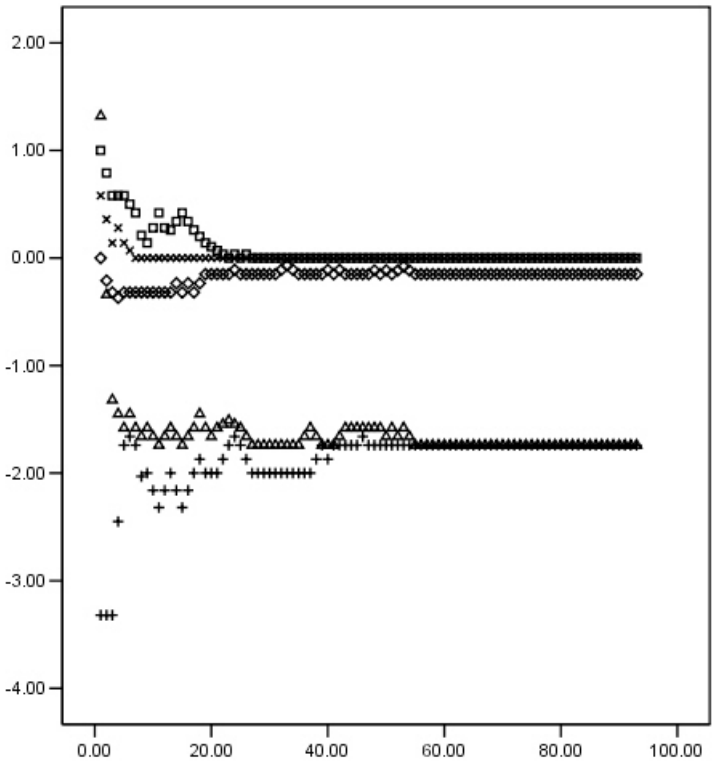
two sessions were also excluded.[92]

This left a total of 85 valid sessions: 48 from the initial recruitment tranche among acquaintances, and 37 from the second group reached via discussion forums. Diathesis variant sentences received a median of 11 responses each. The practise sentences received much higher numbers of responses, and as can be seen from figures 3.1 and 3.2 around 10 or 20 responses were needed to arrive at a stable aggregate measure for each stimulus.

A full table of sentence by sentence results is given in the appendices (see tables C.1 and C.2 on pages 204-205). Each line lists a test sentence in its original (prepositional object) form, followed by its altered (double object) form in brackets. The animacy of the indirect object is indicated in the column headed "Anim.". The original (Orig.) and altered (Alt.) columns give the aggregate magnitude estimation score (the median score across participant responses) given to each variant. The scores are log normalised and are relative to the reference sentence *I can't go into a movie that's*

[92]Excluding these two sessions reduced the data spread considerably: standard deviation for the set of original items dropped from 1.43 to 0.53, and for altered items from 1.58 to 1.13.

Figure 3.2: Convergence of median judgement for practice sentences, as responses were received



The X axis represents the number of responses, the Y axis the median score up to that point.
Symbols represent the same sentence stimuli as in figure 3.1.

Table 3.2: Effect of animacy on success of English dative/benefactive diatheses

	n	**Mean**	**S.D.**	**C.I.**
Animate indirect object	12	-0.16	0.29	±0.16
Inanimate indirect object	12	-0.91	0.61	±0.35

Mean is the mean difference score (i.e. diathesis success) over individual sentences
S.D. is the standard deviation
C.I. is the 95% confidence interval

already started. A score of 0 represents an item of equal acceptability to this reference; +1 twice as acceptable and -1 half as acceptable.

While the scores for altered variants give a measure of the absolute acceptability of each sentence in the double object form, this may be affected by other factors in the dialogue excerpt. As the original sentence is subject to the same factors, the difference between between original and altered variant scores gives an index of the success of the diathesis – the diathesis difference score (the column is headed "Diff."). The columns headed IQR give the interquartile range found among participant responses to original and altered variants. A low value indicates that their was less difference of opinion between judges.

Overall, altered versions performed worse than originals, though in two cases this pattern was reversed. But it is not the case that the double object form is generally less acceptable than the prepositional form, as the results for related fillers demonstrate (table C.3, page 205). In this set, original double object variants out-perform altered prepositional object variants.

The principle hypothesis is examined in table 3.2. The group means of difference scores for sentences with an animate indirect object are compared to those with an inanimate indirect object. The effect of animacy is significant ($p < 0.001$, two-tailed Welch adjusted t-test over sentence difference scores).[93] Thus the principle hypothesis was confirmed.

The next table (3.3) shows the relative effects and significance of the other coded features described in section 3.2.1.5. The table lists the size (n) of the set of sentences that satisfy this feature, out of a total of 24.

The divergence measure "Divg." is a measure of the effect of each feature on the success of the diathesis – it is the difference between the average diathesis success for sentences in the set, and the average for sentences in its complement set (3.100):

(3.100) $divergence = (\overline{o}_f - \overline{a}_f) - (\overline{o}_{\neg f} - \overline{a}_{\neg f})$; where $\overline{o}$ and $\overline{a}$ are means of median aggregate scores of original and altered items, that have the feature f or do not $\neg f$

For example, the set of 14 sentences with an aware indirect object (such as 3.101) generally underwent the dative and benefactive successfully: the mean of original scores ($\overline{o}_f$) over this set was -0.19 and for altered scores ($\overline{a}_f$) was -0.33; so the average difference score for the set is -0.14. For the complement set of sentences with unaware indirect objects (e.g. 3.102) the average difference score was much larger, at -1.45 ($\overline{o}_{\neg f} = -0.05$ and $\overline{a}_{\neg f} = -1.50$), meaning that for these stimuli, the diathesis was much less successful. The divergence score of 1.32 is the difference between these two aggregate success scores ($-0.14 + 1.45$, with rounding), and this measure quantifies the effect of indirect object awareness.

[93] This compares the arithmetic mean for the animate group against that for the inanimate group, over each sentence's difference score (which is itself is a difference between two median scores over participants' responses). The sample sizes are 12 for each group, so the degrees of freedom are 22, or 15.9 with the Welch correction for unequal variances. The normality of the sampling distribution was determined to be adequate for these sample sizes (see page 219 in the appendices).

Table 3.3: Relative contribution of semantic and discourse features to the English dative and benefactive diatheses

Feature	n	Divg.	p	r_{pb}	r^2_{pb}
Aware Indirect Object	14	**1.32**	0.00	0.74	0.55
Affected Indirect Object	12	**1.05**	0.01	0.61	0.37
Positive for Indirect Object	14	**1.01**	0.02	0.57	0.32
Volitional Indirect Object	9	**0.94**	0.01	0.55	0.30
New Direct Object	16	**0.93**	0.06	0.43	0.18
Pronoun or Proper Name Indirect Object	13	**0.80**	0.04	0.46	0.21
Non-Durative Event	13	**0.74**	0.06	0.39	0.15
Given Indirect Object	10	**0.73**	0.06	0.42	0.17
Unaware Direct Object	12	**0.72**	0.07	0.38	0.14

(3.101) a. Is vengeance going to bring your son back to you?[94] (0)

b. Is vengeance going to bring you your son back? (0)

(3.102) a. What in heaven's name brought you to Casablanca?[95] (0.32)

b. What in heaven's name brought Casablanca you? (−1.45)

Only features whose effect was significant ($p < 0.05$) or approached significance ($p < 0.1$) are listed in table 3.3. Features whose samples size fell below that needed to ensure normality of the sampling distribution (the minimum sample size for normality in this sample set was estimated at seven – see figure C.3 on page 219 of the appendices) were also excluded, as the accuracy of the p-values would be compromised. Finally, if two closely related features both had a significant effect, the stronger one was retained and the weaker one was discarded (for example if both the narrow material interpretation of telicity and the broad psychological one had significant effects).

On the basis of the two average difference scores that contribute to the divergence score for a feature, a t-test can estimate how likely this effect is to have come about by chance sampling (in this case, the particular 24 sentences that were chosen as experimental materials). The p-values are two-tailed, calculated using the Welch adjustment to the t-test to allow for unequal sample sizes and for unequal variances (which is often the case, as wellformed sentences tend to receive a lower variance in judgements than illformed ones).[96]

The small sentence samples used prevent the use of multiple linear regression or other techniques to identify interactions between features (for example, the extent to which the effect of an aware and an animate indirect object are independent). The column in the table headed "r" contains the point-biserial coefficient of correlation[97], which is an alternative measure of the magnitude, this time the strength the association between the feature in question and success in a diathesis. The square of this figure, the coefficient of determination estimates what proportion of the variation seen among

[94] *Godfather*, Francis Ford Copola, 1972

[95] *Casablanca*, Michael Curtiz, 1942

[96] Care should be taken when interpreting the p-values of posterior analyses. By the laws of chance, some apparently determining, but in reality spurious patterns will emerge if analyses of enough factors are undertaken. These p-values are indicative of the *relative* likelihood of feature effects to be spurious. Using an α-level of 0.05 can be expected to turn up one false positive result for every 19 true positives. See Sterne and Smith (2001); Ioannidis (2005a,b).

[97] The point-biserial correlation coefficient is a measure of association between a continuous variable x and a binary variable which has values 0 and 1: $r_{pb} = \frac{(\bar{x}_0 - \bar{x}_1)\sqrt{pq}}{s_x}$, where p and q are proportion of 0 and 1 values (so $p = 1 - q$). pq can also be expressed as $\sqrt{\frac{n_0 n_1}{n(n-1)}}$, $n = n_0 + n_1$

Table 3.4: Associations between determining factors of the Dative/Benefactive diatheses

	PIO	**AnIO**	**NDE**	**NDO**	**GIO**	**AfIO**	**VIO**	**UDO**	**AwIO**	**PPIO**
Positive for Indirect Object	-	0.51	0.07	0.30	0.37	0.34	0.65	0.17	0.66	0.41
Animate Indirect Object	0.51	-	0.08	0.18	0.51	0.17	0.60	0.00	**0.85**	0.59
Non-Durative Event	0.07	0.08	-	0.24	-0.07	0.25	0.02	0.25	0.24	-0.01
New Direct Object	0.30	0.18	0.24	-	0.42	0.53	0.18	0.53	0.30	0.24
Given Indirect Object	0.37	0.51	-0.07	0.42	-	0.34	0.57	0.00	0.37	0.61
Affected Indirect Object	0.34	0.17	0.25	0.53	0.34	-	0.26	0.67	0.34	0.08
Volitional Indirect Object	0.65	0.60	0.02	0.18	0.57	0.26	-	-0.09	0.65	0.71
Unaware Direct Object	0.17	0.00	0.25	0.53	0.00	0.67	-0.09	-	0.17	-0.08
Aware Indirect Object	0.66	**0.85**	0.24	0.30	0.37	0.34	0.65	0.17	-	0.58
Pron./Prop. Name Ind. Obj.	0.41	0.59	-0.01	0.24	0.61	0.08	0.71	-0.08	0.58	-

Figures are correlations between pairs of binary features.
Column headings are abbreviations of row headings, following the same order.

the sentences is accounted for by this single feature, *assuming* independence of all of the features considered (which is an unrealistic assumption).

To establish the extent to which these features are confounded in these materials, a confusion matrix was constructed. Table 3.4 shows the strength of association between features, measured using the coefficient of correlation between the binary values of pairs of features.[98] The row headings are repeated in abbreviated form in the column headings. The strongest association seen among these features is between animacy and awareness in the indirect object.

Some of the other strong associations seem reasonable, such that between awareness, volition and positive consequences for the indirect object. The association between lack of awareness in the direct object and affectedness in the indirect object may rather be a product of chance.

3.2.2.5 Discussion

The results for individual sentences seem to support the argument that the admissibility of these diatheses is not solely determined by the verb lexical entry. The best benefactive performer uses the verb *make* (*make ... sausage for us*), as does the second worst (*making statues for ... years*) (see table on page 204 of the appendices). The verb *bring* appears in the second most successful dative diathesis (*bring you your son back*), and the worst (*... brought Casablanca you* – see the continuation of the table on page 205). Similarly, semantic class alone cannot predict the success of the benefactive diathesis. Sentences containing semantically similar verbs are variously productive: both the *fetch* and *catch* sentences work well in the double-object form, while the *find* sentence is unsuccessful in the diathesis (i.e. the original prepositional object form is much preferred).

Evidence on the suggestion that the dative double object construction necessarily involves possession is mixed. The doctor is not a literal recipient in the sentence *I should take a witch doctor my daughter* and this diathesis form is unsuccessful (-0.87) (a mother is refusing to bring her daughter to visit an exorcist). But neither is *you* a recipient in *Is vengeance going to bring you your son back*, which received a neutral difference score of 0.[99]

[98] The correlation between two binary variables, each of which has values 0 and 1, is: $r_\phi = \frac{n_{00}n_{11} - n_{01}n_{10}}{\sqrt{(n_{00}+n_{10})(n_{01}+n_{11})(n_{00}+n_{01})(n_{10}+n_{11})}}$, where n_{ab} is the number of data points that have value a for variable X and value b for variable Y

[99] Levin (1993) lists *take* as a verb that does not admit the dative diathesis. An informal survey among colleagues suggested that that it is as acceptable as *bring*, which Levin does include in dative diathesis verbs. This may be a matter of dialect, as the colleagues in question were Irish, as were many of the experiment participants.

Qualitatively, animacy does appear to be a strong determining factor. Of the top 12 performers, 9 have strictly animate referents in their indirect objects. Sentences with indirect objects that might be considered as animate by holonymic extension give mixed results: *my house, the main office*, and *Gryffindor* (a sports team) perform rather well, while *Work, that video arcade* and *Casablanca* do not. This difference may reflect the difference in the plausibility of these referents as recipients or benefactors, indicating that something more complex that selectional restrictions is at work.

The scores seen here can also be compared to the two practise sentences that contained clear syntactic errors. As can be seen from figure 3.1 (page 96), the two sentences (3.97) and (3.98) receive mean scores of about -1.8 and -2.0 across all participants, and these are similar in magnitude to the three worst altered variants ($-1.78, -1.42, -1.40$) among the benefactive and dative shift candidates. A possible defence of the position that syntactic constraints are inherently "harder" than semantic ones (as is claimed by Chomsky, and Sorace and Keller, 2005 among others), would be that animacy is recorded in the lexicon as a syntactic feature, and that the lack of this feature in *Casablanca*, *bread*, and *statues* causes the unacceptability of the worst dative and benefactive shifts. However phrases can also be found whose head's lexical entries would be expected to be inanimate (e.g. *main office* or *Gryffindor*), but nevertheless are successful as the indirect object of double object constructions. The sentences containing these arguments each received scores close to 0 (i.e. close to the reference) for their altered variants, indicating that acceptability is indeed dependent on interpretation, rather than just structure.

The patterns seen in the table of feature effects are consistent with many expectations supplied by the literature. Interestingly, the numbers suggest that the relational notion of awareness (divergence score 1.32) in the indirect object is a better predictor of the dative and benefactive double object construction, than the ontological feature of animacy (divergence score 0.75), though since these two features are confounded the difference may be spurious. Positive outcome and volitional involvement of the indirect object are consistent with psychological affectedness, and with traditional role accounts built around Beneficiaries and Recipients. Features of perspective also seem to play a role. An animate indirect object would fall above an unaware direct object on a scale of anthropocentric saliency, promoting its placement towards the start of the sentence, and this is also consistent with an indirect object that is given, and a direct object that is new to the discourse. The effect of non-durativity is hard to explain, and does not receive support from the literature.

3.2.3 The English Passive Diathesis

3.2.3.1 Background

The foregrounding and backgrounding of participants is one of the principle functions of the passive diathesis (see section 3.2.1.2). If accounts of a "natural topic hierarchy" (e.g. Croft, 1991) are correct, then there will be a general tendency to put animate participants in the prominent subject position. Animacy is also a prerequisite for the role of Experiencer, which also is associated with the subject (Dowty, 1991).

The primary hypothesis to be tested is whether the animacy of the direct object in active sentences is a determining factor in the success of the passive diathesis. Authentic sentences which have the configuration NounPhrase_1–Verb–NounPhrase_2 will be compared to their NounPhrase_2–BE–VerbED–by–NounPhrase_1 variant. The difference between the acceptability scores of the original and altered variants will give a measure of diathesis success, and group means of sentence diathesis success will be compared to determine if animacy of the canonical object has a significant

effect.

Other factors that have been suggested as determinants in the success of the passive diathesis include: the agency of the (active canonical) subject; affectedness of the direct object (Jackendoff, 1972; Palmer, 1994); animacy or awareness in the subject; the stative or dynamic nature of the predicate; and the relative heaviness of subject and object (see section 3.2.1.2). As before, post-hoc comparisons will be made of the effect on the passive of semantic and pragmatic features described in section 3.2.1.5, and the results of individual sentences will be examined.

3.2.3.2 Materials

The passive/active experimental items were selected together with other English test and filler items as discussed in section 3.2.2.2. Simple finite declarative active sentences such as (3.103) were chosen with an overt subject (e.g. no non-subject imperatives, and no expletive *it* or *there*), a single bare noun phrase object, and no gapped arguments (for example in relative clauses), using any verb other than *be* or *have*.[100]

(3.103) a. **Man at Elevator**: Someone$_{subj}$ saw a cockroach$_{dobj}$ up on twelve.[101]

b. **Man at Elevator**: A cockroach$_{subj}$ was seen by someone$_{oblq}$ up on twelve.

As the objective of this experiment was to compare syntactic and semantic constraints, the structural requirements of the diathesis were interpreted very literally. For example a bare noun phrase such as *home* was treated as a direct object (3.104), despite the fact that it might be considered a spatial adverbial phrase (as it could be replaced by a phrase such as *back to the house*).[102]

(3.104) a. **T-1000 impersonating Janelle**: ... If you$_{subj}$ hurry home$_{dobj}$ we can have dinner together.[103]

b. **T-1000 impersonating Janelle**: ... If home$_{dobj}$ is hurried by you$_{oblq}$ we can have dinner together.

The composition of the materials were as described table 3.1 in the section on the dative/benefactive diathesis (page 93). The test sentences were evenly split by animacy – 15 of the original active/passive sentences had a direct object that was animate, and 15 inanimate.[104]

3.2.3.3 Procedure

The English passive stimuli were presented in the same experiment as the dative/benefactive stimuli. See section 3.2.2.3.

[100] As discussed in the introduction to the passive construction, English has both stative and dynamic passives. The passives examined here are by definition dynamic, since they are derived from active counterparts.

[101] *Ghostbusters*, Ivan Reitman, 1984.

[102] In addition to (3.104), one other active sentence was selected, whose bare noun phrase argument would not be viewed as a direct object in some syntactic frameworks: *Crucifixion lasts **hours***. As will be seen, these two sentences were the worst performers in the passive, which is consistent with the view that these are syntactically or semantically different.

[103] *Terminator 2: Judgement Day*, James Cameron, 1991.

[104] As mentioned in section 3.2.1.5, a classification by aspectual types was attempted, but proved unsatisfactory. However, the provisional classification by aspectual types was: 6 Process items, 6 Culminated Processes, 6 Culminations, 6 States and 6 Points (though many of these were marginal cases), using the terminology of Moens and Steedman (1988).

Table 3.5: Effect of animacy on success of English passive diathesis

	n	Mean	S.D.	C.I.
Animate direct object	15	-0.34	0.39	±0.20
Inanimate direct object	15	-0.58	0.45	±0.23

Mean is the mean difference score (i.e. diathesis success) over individual sentences
S.D. is the standard deviation
C.I. is the 95% confidence interval

Table 3.6: Relative contribution of semantic and discourse features to the English passive

Feature	*n*	Divg.	*p*	r_{pb}	r^2_{pb}
Positive Event	16	**0.49**	0.08	0.32	0.10
New Subject	8	**0.45**	0.10	0.30	0.09

3.2.3.4 Results

The analyses of the passive diathesis responses were carried out in a manner identical to that used for the dative/benefactive data. As with the dative and benefactive sentences, altered passive variants were generally less acceptable than their original active counterparts (see table C.5 in the appendices, page 206), though as the passive sentences among the related fillers in table C.3 (page 205) demonstrate, passives are not universally dispreferred. A glance at the ranking of active/passive items by difference scores reveals that the animacy of the canonical direct object does not seem to be a major determining factor in judged acceptability.

Table 3.5 shows the mean difference scores (i.e. measures of success) for sentences, depending on the animacy of the canonical (i.e. active) direct object (which appears as the subject in the passive form). The effect was only slight, approaching significance at $p = 0.07$.[105]

Only two factors emerged in the posterior analysis of features are given in table 3.6. Neither achieved significance.[106] The effect of positive affect (if it is not spurious) is hard to explain, as any suggestions of its action have seen *negative* affect as promoting the passive. The effect of the discourse status of the canonical subject (i.e. *by* prepositional object in the passive) does make sense, as there is a general preference to put new entities towards the end of the clause. None of the features were strongly confounded (table 3.7).

[105] One-tailed *t*-test of independent means over sentence difference scores, with Welch correction for unequal variances, $n = 30$, $df = 27.34$. This effect was less significant $p = 0.12$ if the sentence (3.104) is excluded.
[106] Both are less significant if sentence (3.104) is excluded: $p = 0.14$ and $p = 0.15$.

Table 3.7: Associations between determining factors of the English passive diathesis

	PE	ADO	NS
Positive Event	-	0.15	0.15
Animate Direct Object	0.15	-	0.30
New Subject	0.15	0.30	-

Figures are correlations between pairs of binary features.
Column headings are abbreviations of row headings, following the same order.

3.2.3.5 Discussion

None of the features examined achieved significance. Of those that approached significance, the animacy of the canonical direct object and the newness of the canonical subject, seem reasonable for pragmatic reasons. According to notions of anthropocentric saliency (Fillmore, 1977b; Croft, 1991) speakers are more likely to put animate entities towards the start of a sentence (the canonical direct object becomes the subject in the passive), and for information structuring reasons new elements tend to be postponed. The positive affect of the event is hard to explain.

Verbs and verb classes alone did not appear to be a strong predictor either, as similar or identical verbs can be found at opposite ends of the ranking. The sentences *I was loved by her* and *The British are hated by me* are similar in sense, but have difference scores of 0 and -1 respectively. The two sentences using the verb *hit* also vary in the success of the diathesis (*he was hit by me:* -0.42; *the wall is hit by you*: -1.00), but these use different senses of the word.

The worst performers were derived from active sentences whose post-verb noun phrase might not be considered an object – the Measure phrase in *Crucifixion lasts* ***hours***, and the Goal phrase in *you hurry* ***home***. While these could be used to argue that structural constraints produce larger acceptability effects, both of these cases have semantic justification also, since neither is animate, or is Patient in any way. In addition, the first sentence is stative.

3.2.4 The German Passive Diathesis

3.2.4.1 Background

The motivation for this experiment is similar to that on the English passive (section 3.2.3). According to accounts of a "natural topic hierarchy" or anthropocentric saliency, speakers will prefer to place animate participants in the subject position.

The primary hypothesis to be tested is whether the animacy of the direct object in active sentences is a determining factor in the success of the passive diathesis. Authentic sentences which have the configuration NounPhrase_1–Verb–NounPhrase_2 will be compared to their *Vorgangspassiv* (dynamic passive) counterparts of the form NounPhrase_2–WERDEN–GEVerbT–VON/DURCH–NounPhrase_1. The difference between the acceptability scores of the original and altered variants will give a measure of diathesis success, and group means of sentences diathesis success will be compared to determine if animacy of the prepositional object has a significant effect.

Other factors that have been suggested as determinants in the success of the passive diathesis include the transitivity of the event described (in contrast to stative events); volition and causation on the part of the canonical subject; and foregrounding of the canonical object (as discussed in section 3.2.1.3). Post-hoc comparisons will also be made of the effect on the passive of semantic and pragmatic features described in section 3.2.1.5, and the results of individual sentences will be examined.

3.2.4.2 Materials

For the German materials fewer fan-contributed film quotes could be found. As a result, the English materials were translated (this was performed by a colleague at the Trinity College Dublin who is a native speaker of German and a trained translator[107]), and then supplemented with a smaller

[107] The author thanks Conny Opitz, of the Department of Russian. See Murphy et al. (2005).

number of sentences from original German cinema. Some of the translations were somewhat problematic, since the aim was to have both a felicitous translation, and a similar syntax (Subject–Verb–DirectObject) so that the German sentence would also have the appropriate structure to make it a candidate for the passive. In such cases the most felicitous translation available that kept the same entities in subject and non-subject grammatical functions was taken. For example in (3.105) the English direct object *captain* is realised as an oblique *vorm Captain* 'in front of the captain' in German. In the next example, the German requires a third dative object *dir* 'to you' to be added, and the English *it* must be realised as *Chance*.

(3.105) ... wenn du_{subj} vorm $Captain_{oblq}$ salutierst ...
... when you before-the captain salute ...
'... when you_{subj} salute the $captain_{dobj}$ ...'

(3.106) du_{subj} hast dir_{iobj} die $Chance_{dobj}$ vergeben
you have to-you the chance wasted
'you_{subj} wasted it_{dobj}'[108]

Ten German language films were used to gather materials. The dialogues came from a variety of sources, including a collection of full scripts[109], one website of enthusiast submitted film quotes[110], and collections of DVD subtitle files on the internet.[111] An interesting complication that this introduced was that subtitles are sometimes substantially different from the dialogue they accompany, usually shortened for reasons of screen-space – compare the original spoken dialogue (3.107a), the subtitles it is distributed with on DVD (3.107b), and the form as posted online by a web-user (3.107c). Nevertheless, all are wellformed and constitute authentic language use.

(3.107) a. Machen Sie Photos von einlaufende Besatzungen, nicht von auslaufende ...
make you photos of in-walking crews, not of out-walking ...
'Photograph the crew going on board, not coming off ...[112]

b. Fotografieren Sie lieber, wenn wir wieder einlaufen ...
photograph you rather, when we again in-walk ...
'You would be better photographing us again when we are going on board ...

c. Filmen sie einlaufende Besatzungen, nie auslaufende ...
film you in-walking crews, never out-walking ...
'Film crews going on board, never coming off ...

The diathesis of interest is the passive, so simple active sentences based around a monotransitive verb with noun-phrase subject and noun-phrase direct object were chosen (3.108a) and manually altered to yield a passive variant (3.108b). Again, the structural criteria were applied strictly, without any appeal to theories of covert syntax, meaning that measure (3.108) and predicative objects (*heißt das* ***Kosmonaut*** 'it's called cosmonaut') were included in the experimental set. However, as in the English passive materials, these were small in number – the only examples are the two sentences just mentioned.

[108] *Rocky*, John G. Avildsen, 1976.
[109] [http://www.zweitausendeins.de/filminfo/Boerse.htm], viewed August 2006.
[110] [http://filmzitate.ichel.de/], viewed August 2006.
[111] These subtitle sites appear to be illicit, as they can be used when making copies of DVDs. The websites seem to move quite regularly, but such sites can be found by searching with the keyword "subtitles".
[112] *Das Boot*, Wolfgang Petersen, 1981.

(3.108) a. Die DDR_{subj} wurde 40_{dobj}
the GDR became 40
'The German Democratic Republic was 40 years old'[113]

b. *40_{subj} wurde durch die DDR_{oblq} geworden
40 was through the GDR became
'40 was become by the GDR'

The canonical subject was introduced either with *von* 'by' (3.109a) or *durch* 'through, by means of' (3.110a), depending on the degree of agentivity (Dodd and Zojer, 2003, p.104) (these two sentences are related filler sentences, which were originally in the passive form).

(3.109) a. Die meisten Kojen werden von 2 Männer benutzt
the most bunks become by 2 men used
'Most of the bunks are used by two men'[114]

b. 2 Männer benutzten die meisten Kojen
2 men use the most bunks
'2 men use most of the bunks'

(3.110) a. ... wenn das Feuer durch Pfusch bei der Installation verursacht wurde ...
... if the fire through botch at the installation caused became ...
'... if that fire was caused by kooky wiring ...'[115]

b. ... wenn Pfusch bei der Installation das Feuer verursacht hat ...
... if botch at the installation the fire caused has ...
'... if kooky wiring caused that fire ...'

Sentences were excluded for the same reasons as in English (see section 3.2.2.2), with the additional limitation that the subject preceded the direct object (as mentioned on page 78, German allows object fronting). Thus sentences like (3.111) were disallowed. Twenty valid authentic German active sentences were gathered.

(3.111) Dem_{dobj} folgt meistens das große $Schweigen_{subj}$, Stille.
that follows mostly the big silence, quiet
'That is followed by the great silence, quiet' or 'The great silence, quiet, follows that'[116]

The same six reference and practice sentences were used in translation from English (3.112-3.117), with the same two sentences being purposefully manipulated to make them illformed (3.116-3.117).

(3.112) ... ich kann nicht in einen Film gehen, der schon angefangen hat
... I can not in a film go, that already started has
'I can't go into a movie that's already started'

(3.113) das kann etwas länger dauern
... that can somewhat longer last
'... it may take a little longer'

(3.114) Hatte mir jemand 'ne Käsepizza mitbestellt
had to-me someone a cheese-pizza ordered-as-well
'Did anyone order me a plain cheese'

[113] *Good Bye Lenin!*, Wolfgang Becker, 2003.
[114] *Das Boot*, Wolfgang Petersen, 1981.
[115] *The Towering Inferno, Das Boot*, Wolfgang Petersen, 1981.
[116] *Das merkwürdige Verhalten geschlechtsreifer Großstädter zur Paarungszeit*, Marc Rothemund, 1998.

Table 3.8: Summary of experimental materials for German acceptability experiment

	Original	Translated	Altered	Total
General Fillers	8	4	-	12
Related Fillers	6	8	14	28
Reference/Practise	-	4	2	6
Passive Test	20	30	50	100

(3.115) ... [ich] war im Gegenteil von tiefer Trauer erfüllt
... [I] was in-the contrary from deep sadness filled
'I was surrounded by an endless sorrow'

(3.116) Wirf das Götzenbild mir
throw the idol to-me
'Throw the idol me '

(3.117) Hör endlich immer nur an Sex zu denken auf
give at-last always only on sex to think up
'Stop thinking on sex about '

The composition of materials is shown in table 3.8. The total number of dialogue extracts was 82, and including diathesis altered variants, the number of sentence stimuli was 146. The active/passive test sentences were evenly split by the animacy of the canonical direct object.[117]

3.2.4.3 Procedure

The procedure used was identical to that used in the English-language experiments described in sections 3.2.2 and 3.2.3. The instructions were translated from the English original, and are included in the appendices (see page 187). A full pilot of the web-based experiment was carried out among 14 acquaintances, to validate the methodology, software, materials and instructions. Some minor errors in materials were discovered during piloting, and the instructions underwent revisions in response to pilot feedback.

The experiment was then administered over the web to respondents who were recruited by email (all direct or indirect acquaintances of the experimenters). Participants were asked to score 30 sentences, presented three per page, again with full context (surrounding dialogue, scene description and film description).

3.2.4.4 Results

More responses were received for the German experiment, than for the other two languages. Of a total 243 sessions, 62 were discarded because the respondent did not start the experiment proper, or dropped out within the warm-up stage of practise items, and 7 respondents were excluded for backtracking to change responses. If respondents gave a lower average response to the two intentionally unacceptable practise sentences (3.116-3.117), than they did to the three attested and acceptable sentences (3.113-3.115), they were excluded for having inverted the judgement scale (2 sessions).

[117] As mentioned in section 3.2.1.5, a classification by aspectual types was attempted, but proved unsatisfactory. However, the provisional classification by aspectual types was: State 8, Point 2, Process 2, Culminated Process 2 and Culmination 6, using the terminology of Moens and Steedman (1988).

Table 3.9: Effect of animacy on success of German passive by means of sentence difference scores

	n	Mean	S.D.	C.I.
Animate direct object	25	-0.60	0.63	±0.25
Inanimate direct object	25	-1.08	0.70	±0.27

Mean is the mean difference score (i.e. diathesis success) over individual sentences
S.D. is the standard deviation
C.I. is the 95% confidence interval

Nineteen sessions were excluded because the respondents were non-native speakers of German. No facetious or obviously mistaken responses were found.

The highest magnitude estimate given was of 20 times the reference (for the sentence *Ich verstehe überhaupt nichts mehr* 'I don't get it anymore'), and the lowest a score of 1/100 (for the sentence *40 wurde durch die DDR geworden* '40 was become by the GDR'). Though these values were extreme, examination of the sessions logs indicated that they were not facetious.

A total 153 valid sessions remained for analysis, and diathesis variants sentences received a median 33 judgements each (again, fillers and practise sentences received higher numbers). Since the German materials received more responses than those of the other two languages, they were examined to see if there was an order effect (as noted in section 3.1.2, participants may become less careful as the experiment progresses, as they become bored or habituated). The full set of responses were examined to see if scores given were systematically higher or lower as each session progressed, but though the linear correlation found was significant, it was very slight ($r = 0.046,\ p = 0.035$). It was also not the case that responses became considerably more noisy – divergence of individual scores from each item's final aggregate score did not change as the session progressed ($r = -0.06,\ p = 0.063$).

The full listing of sentence by sentence results for the German test materials is given in tables C.6, C.7 and C.8 (on pages 207 to 209 of the appendices). The table for related fillers follows on page 210.

The negative judgements given by German respondents were stronger than those seen in the English experiments. Again, in general, original active sentences were better than altered passive sentences among the test materials, though the fact many original passives were similarly or more acceptable to altered actives suggests that the passive form is not universally dispreferred. German original active sentences were judged marginally superior to those that were translated from English, but the difference was not significant (source language is indicated in the sentence tables using the codes of 'en' for English and 'de' for German).[118] However in individual cases, this may account for translated items coming towards the top of the list (if a translated original score is lower than normal, this can result in a higher difference score, even when the passive variant is of marginal acceptability).

Animate direct objects (of the active sentence, that appear as subject in the passive) resulted in more successful passive sentences than inanimate objects (see table 3.9), to a much stronger degree to that seen in English ($p = 0.007$, two-tailed *t*-test of independent samples over sentence difference scores with Welch adjustment).

The posterior analysis of additional features (table 3.1) shows that effects in German are also stronger than those seen for the English passive. The smaller *p*-values are a product both of the larger effect magnitudes, and the larger sample sizes (there were 50 German test items for the passive, but

[118] $p < 0.2$, two-tailed *t*-test of independent samples over sentence scores with Welch correction, n=80.

Table 3.10: Relative contribution of semantic and discourse features to the German passive

Feature	n	Divg.	p	r_{pb}	r^2_{pb}
Materially Telic Event	16	**0.82**	<0.01	0.39	0.16
Negatively Affected Object	16	**0.78**	<0.01	0.41	0.16
Materially Affected Object	14	**0.77**	0.01	0.39	0.15
Causing Subject	33	**0.71**	0.02	0.33	0.11
New Subject	11	**0.64**	0.05	0.30	0.09
Aware Object	21	**0.53**	0.05	0.26	0.07
Given Object	18	**0.50**	0.06	0.27	0.08

Table 3.11: Associations between determining factors of the German passive diathesis

	NAfO	MTE	NS	GO	CS	MAfO	AwO
Negatively Affected Object	-	0.36	0.26	0.38	0.31	0.43	0.20
Materially Telic Event	0.36	-	0.26	0.20	0.31	**0.81**	0.02
New Subject	0.26	0.26	-	0.21	0.18	0.31	0.13
Given Object	0.38	0.20	0.21	-	0.36	0.27	0.29
Causing Subject	0.31	0.31	0.18	0.36	-	0.35	0.27
Materially Affected Object	0.43	**0.81**	0.31	0.27	0.35	-	0.10
Aware Object	0.20	0.02	0.13	0.29	0.27	0.10	-

Figures are correlations between pairs of binary features.
Column headings are abbreviations of row headings, following the same order.

only 30 in English).

Telicity of the event and affectedness in the canonical object contribute strongly to the formation of the passive, though these two features are so strongly associated that their effects cannot be considered independent (they have a mutual binary correlation of 0.81 – see the confusion matrix in table 3.11). Causation by the canonical subject is another feature of transitivity that proves to be a factor. Negative affect for the canonical object appears to have an effect that is independent of material affectedness (mutual correlations are low). Given, aware objects are eager to take the position of subject in the passive. And new subjects are eager to move after the verb.

3.2.4.5 Discussion

There are few clear examples of verbs, or verb classes that vary between successful and unsuccessful passive formation. There are three sentences using the verb *brauchen* 'need', with varying difference scores (-0.42, -1.00, -1.42), but none of these can be seen as successful.

Again, active sentences which did not have a transitive object were least successful in the passive – verbs with predicate (*heißen* 'be called'; *werden* 'become'), extent (*dauern* 'last') or location (*kommen* 'come') objects. Possible explanations include differences in syntax, the stative aspect of some of the verbs, or the lack of affectedness in their objects.

Generally, transitive properties (telicity, affectedness of the object, causation by the subject) contribute to the success of the passive in German. Interestingly, negative affect appears to have an effect that is independent of material affectedness, suggesting that the German passive has a similar adversative function to that of the Chinese *bei* construction.

Animacy and givenness of the direct object, and newness in the subject of an active original sentence promoted the diathesis, showing that discourse factors operate in the German passive also.

Since 30 test sentences were shared between the English experiment and the German experiment (in translation), and the fact that they appear to share many of the same tendencies in the formation of the passive, it was instructive to establish if there is a strong correlation of passive acceptability across languages. Indeed a high Pearson correlation of 0.7 ($p < 0.01$) was found between the individual difference scores for sentences that were used in both the English and German experiments.

3.2.5 The Chinese *bei* and *ba* Diatheses

3.2.5.1 Background

The motivation for this experiment is similar to that on the English and German passives (sections 3.2.3 and 3.2.4). Anthropocentric saliency may cause speakers to place constituents that refer to animate participants closer to the head of the sentence, to give them prominence. As discussed in section 3.2.1.4, the *bei* construction reverses the canonical order, putting the object before the canonical subject. The *ba* construction promotes the object to a position before the verb, but still following the subject (canonical order in Chinese is Subject–Verb–Object).

This experiment will test the hypothesis that animacy of the canonical object in sentences such as (3.118a – repeated from page 81) promotes the success of both diatheses. Authentic sentences which have the configuration Subject–Verb–NounPhrase–Resultative (3.118a) will be compared to manually constructed *bei* (3.118b) and *ba* (3.118c) variants. The difference between the acceptability scores of the original and altered variants will give a measure of diathesis success, and group means of sentences diathesis success will be compared to determine if animacy object has a significant effect.

(3.118) a. tamen$_{\text{subj}}$ qieduan le dianli$_{\text{dobj}}$
they cut PFV electric-power
'They cut the power'
它们切断了电力

b. dianli$_{\text{dobj,topic}}$ bei tamen$_{\text{oblq}}$ qieduan le
electric-power BEI they cut PFV
电力被它们切断了

c. tamen$_{\text{subj}}$ ba dianli$_{\text{oblq}}$ qieduan le
they BA electric-power cut PFV
它们把电力切断了

The post-hoc investigation of additional features (described in section 3.2.1.5) will investigate claims of what other factors influence the success of these two constructions. Transitivity and affectedness, and discourse structuring functions have been ascribed to both (see section 3.2.1.4). Individually, the *bei* construction is said to have an adversative function, and the *ba* construction may be associated with concrete (rather than abstract) events. Features will be compared by the strength of their effect on diathesis success.

3.2.5.2 Materials

The collection of Chinese materials was handled similarly to those in German. The English materials were translated in their entirety by a professional translator, and supplemented with materials from eleven Chinese films. Two sentences could not be successfully translated into the form of a monotransitive verb phrase with a direct object: ... *salute the captain* ... would have required a

coverb phrase that would have appeared before the verb (*xiang duizhang* 向队长 'to the Captain') and (3.119) would lack an object altogether, as pronominal arguments can be dropped in Chinese.[119]

(3.119) ... wo bu runxu
... me not permit
'I forbid it'[120]
我不容许

The excerpt below from *Infernal Affairs III* shows again variation between the script as spoken on screen, as subtitled and as remembered by fans. The Hong Kong slang term 马夫 *mafu* 'pimp' in the spoken original (3.120) was subtitled as 皮条客 *pitiaoke* 'pimp', presumably to be intelligible to the general Chinese viewer. In the fan's rendition both terms for *pimp* are given, the leading 其实 *qishi* 'actually' is left out and the variant 偷泡 *toupao* 'steal' is substituted for 泡 *pao* 'steal'.

(3.120) qishi wo ye shi mafu, wo kan ta you mei you pao wo de niu
actually me also be pimp, me see him have not have steal me MOD girl
'Actually I'm a pimp too – I'm checking that he hasn't stolen any of my chicks'
其实我也是马夫，我看他有没有泡我的妞

Chinese monotransitive sentences with both arguments present were searched for among the film materials available – something more of a challenge as Chinese often drops arguments (e.g. the subject is absent in 3.121) in cases where a pronoun would be used in English or German. As described in section 3.2.1.4, both the *ba* and *bei* constructions have the same structural requirements for a resultative complement on the verb (such as *chulai* in 3.121), so a sentence such as (3.122) was not admissible. Twenty candidate monotransitive active sentences were gathered.

(3.121) zixi kan neng kan chulai
detail look can look out
'If one looks carefully, one can tell'
仔细看能看出来[121]

(3.122) ni shifu kexi tai xiaokan nüren
you master regrettably too look-down-on woman
'Unfortunately your master looks down on women'
你师父可惜太小看女人[122]

The same six reference and practice sentences were used, in translation (3.123-3.128), the last two being manually altered to render them illformed.

(3.123) wo bu neng jinqu kan yijing kaishi de dianying
I not can go-in see already start ATTR film
'I can't go into a movie that's already started'
我不能进去看已经开始的电影

(3.124) zhe yang hui jiu yi dian
this kind can long a little
'... it may take a little longer'
这样会久一点

[119] See footnotes on page 80.
[120] *Titanic*, James Cameron, 1997.
[121] *In the Mood for Love*, Wong Kar-wai, 2000.
[122] *Crouching Tiger, Hidden Dragon*, Ang Lee, 2000

Table 3.12: Summary of experimental materials for German acceptability experiment

	Original	**Translated**	**Altered**	**Total**
General Fillers	5	5	-	10
Related Fillers	8	8	16	32
Reference/Practise	-	4	2	6
***ba/bei* Test**	20	30	50	100

(3.125) you ren bang wo dian le zhi you ganlao de pisabing ma
there-be person help me order PFV only there-be cheese ATTR pizza INTR
'Did anyone order me a plain cheese'
有人帮我点了只有干酪的披萨饼吗

(3.126) bei yi zhong jimie de beiai huanrao
BEI a kind nirvana ATTR sad surround
'I was surrounded by an endless sorrow'
给我把头像

(3.127) gei wo ba touxiang
give me grip idol
'Throw the idol me'
给我把头像

(3.128) ni you bu yao xiang dao xing le
you again not will think at sex PFV
'Stop thinking sex about'
你又不要想到性了

The experimental set was composed of 50 test sentences for the *ba/bei* diatheses (20 in the original Chinese, 30 translated from English), 10 general fillers (half original, half translated) and 16 related fillers (half original, half translated). As table 3.12 shows, the total number of dialogue extracts was 82, and including diathesis altered variants, the number of sentence stimuli was 148. The *ba/bei* test sentences were split evenly by animacy.[123]

3.2.5.3 Procedure

The procedure used was identical to that used in the English and German-language experiments described in sections 3.2.2-3.2.4. The instructions were translated from the English original, and are included in the appendices (see page 190). A full pilot of the web-based experiment was carried out among 8 acquaintances. Some minor errors in materials were discovered during piloting, and the instructions underwent revisions in response to pilot feedback.

The experiment was then administered over the web to respondents who were initially recruited by email (all direct or indirect acquaintances of the experimenters), and then via web forums. Participants were asked to score 30 sentences, presented three per page, again with full context (surrounding dialogue, scene description and film description).

[123] As mentioned in section 3.2.1.5, a classification by aspectual types was attempted, but proved unsatisfactory. However, the provisional classification by aspectual types was: 3 State items, 3 Points, 4 Processes, 5 Culminated Processes and 5 Culminations, using the terminology of Moens and Steedman (1988).

3.2.5.4 Results

There were many fewer valid responses to the Chinese experiment. Of a total 112 sessions, 60 were discarded because the respondent did not start the experiment proper, or dropped out within the warm-up stage of practise items, and three respondents were excluded for backtracking to change responses. If respondents gave a lower average response to the two intentionally unacceptable practise sentences (3.127-3.128), than they did to the three attested and acceptable sentences (3.124-3.126), they were excluded for having inverted the judgement scale (four sessions). Three sessions were excluded because the respondents were non-native speakers of Chinese. No facetious or obviously mistaken responses were found.

Forty two valid sessions remained to be entered in the analysis. Due to these lower numbers, and the fact that there were two diathesis variants (*ba* and *bei*) for every active sentence, variant sentences received only a median four judgements each. The range of judgements given by Chinese respondents was also generally larger than for the other two experiments. So, given the small number of judgements gathered per test stimulus, the figures for individual sentences should be treated more cautiously.

The full listing of sentence by sentence results for the *ba* construction are given in tables C.10-C.11 and for the *bei* construction in tables C.12-C.13 (on pages 211 to 214 of the appendices). The table for related fillers follows on page 215. The source language of the each sentence is indicated with the labels 'en' for English, and 'zh' for Chinese.

Again, in general, original active sentences were better than altered passive sentences among the test materials, though the fact that many original *bei* and *ba* sentences among the related fillers were similarly or more acceptable to altered actives suggests that these two noncanonical constructions are not universally dispreferred. Among the original sentences, those that were translations from English were less acceptable than those originally in Chinese.[124] This accounts for the top-ranking sentence among both *bei* and *ba* diatheses (*you wasted it*) – both altered variants received reasonable scores (of 0.21 and 0.06 respectively), but the low median magnitude estimate of -0.74 given to the translated active variant resulted in large positive difference scores. It also appears that respondents found the translated reference sentence (3.123) of rather marginal acceptability, since the median acceptability score given to unaltered authentic Chinese sentences was 0.4.

Looking at the tables of *bei* and *ba* sentences, active sentences with an inanimate object appear to have been marginally more successful. Table 3.13 summarises their performance. This effect is opposite to that expected, but weak, and not significant at conventional levels.[125]

The posterior analysis of additional features shows a range of effects. *Bei* (see table 3.14) shows some expected patterns, preferring a telic event that materially effects its object (i.e. a transitive event). The association with concrete events is interesting, as it is something usually attributed to the *ba* construction. None of the accounts reviewed (in section 3.2.1.4) predicted a preference for events of little or no duration. The expected association with negative affect was seen, but only approached significance.

No strong discourse related factors emerged. Active sentences with given objects were marginally more successful, but this effect did not approach significance.

Table 3.15 shows the corresponding figures for the *ba* construction. Again, non-durative transitive events are preferred. Benefit for the subject also seems to be a factor, as are negative con-

[124] $p < 0.001$, two-tailed t-test of independent samples over sentence scores with Welch correction, n=100.

[125] $p = 0.25$ for the *ba* diathesis, and $p = 0.36$ for *bei*, using a two-tailed t-test of independent samples over sentence difference scores, with Welch correction.

Table 3.13: Effect of animacy on success of Chinese *bei* and *ba* constructions by means of sentence difference scores

Diathesis	***n***	**Mean**	**S.D.**	**C.I.**
***ba*, animate direct object**	25	-1.19	0.72	±0.28
***ba*, inanimate direct object**	25	-0.92	0.89	±0.35
***bei*, animate direct object**	25	-1.19	0.93	±0.36
***bei*, inanimate direct object**	25	-0.94	0.99	±0.39

Mean is the mean difference score (i.e. diathesis success) over individual sentences
S.D. is the standard deviation
C.I. is the 95% confidence interval

Table 3.14: Relative contribution of semantic and discourse features to the Chinese *bei* construction

Feature	***n***	**Divg.**	***p***	r_{pb}	r^2_{pb}
Non-Durative Event	26	**0.71**	0.01	0.44	0.20
Telic Event	32	**0.63**	0.03	0.36	0.13
Materially Affected Object	14	**0.57**	0.05	0.34	0.11
Concrete Event	19	**0.55**	0.04	0.36	0.13
Negative Event	20	**0.47**	0.08	0.27	0.07

sequences (particularly for the object – something generally associated with *bei*). The effect of concreteness only approached significance. Givenness of the object and newness of the subject were marginally associated with the construction, but again these effects did not approach significance.

The confusion matrix below shows the strength of pairwise associations between the factors whose effects achieved or approached significance. Telicity is associated with concreteness and lack of duration in the event, but neither of these are strongly confounded.

3.2.5.5 Discussion

Verb classes or verb senses did not appear to be sole licencing factors in the success of either diathesis. Three clauses using the same verb 看到 *kandao* 'see, spot' received median scores of -0.42, -1.01 and -1.53 for their *ba* variants; and 0.08, -0.15 and -0.55 for *bei* variants. The related filler sentences with the semantically similar verbs 插刀 *chadao* 'stab' and 扎伤 *zhashang* 'impale' also performed differently when moving from a *bei* to an active configuration.

Two verb pairs that nicely illustrate the telic nature of the constructions are 爱/爱上 *ai/aishang* 'love/fall in love' and 找/找到 *zhao/zhaodao* 'search/find'. The stative *love* is less successful in both

Table 3.15: Relative contribution of semantic and discourse features to the Chinese *ba* construction

Feature	***n***	**Divg.**	***p***	r_{pb}	r^2_{pb}
Telic Event	32	**1.32**	0.01	0.44	0.19
Positive for Subject	13	**1.09**	0.01	0.39	0.15
Non-Durative Event	26	**1.05**	<0.01	0.43	0.18
Affected Object	23	**1.01**	<0.01	0.42	0.18
Negative for Object	17	**0.94**	0.01	0.38	0.14
Negative Event	20	**0.72**	0.05	0.28	0.08
Concrete Event	19	**0.61**	0.07	0.25	0.06

Table 3.16: Associations between determining factors of the *bei* and *ba* diatheses

	PS	**NE**	**NO**	**TE**	**NDE**	**CE**	**MAfO**	**AfO**	**AwO**
Positive for Subject	-	-0.02	0.34	0.25	0.02	0.19	0.34	0.18	0.09
Negative Event	-0.02	-	0.45	0.19	0.21	0.12	0.22	0.23	0.18
Negative for Object	0.34	0.45	-	0.19	0.18	-0.04	0.40	0.61	-0.10
Telic Event	0.25	0.19	0.19	-	0.53	0.50	0.37	0.27	0.06
Non-Durative Event	0.02	0.21	0.18	0.53	-	0.26	0.24	0.32	-0.08
Concrete Event	0.19	0.12	-0.04	0.50	0.26	-	0.06	0.10	-0.02
Materially Affected Object	0.34	0.22	0.40	0.37	0.24	0.06	-	0.68	-0.05

Figures are correlations between pairs of binary features.
Column headings are abbreviations of row headings, following the same order.

constructions than the telic *fall in love*, while the continuing *search* is worse as a *ba* sentence than the completed *find*.

Overall, the *ba* and *bei* constructions differ considerably from the passive constructions to which the are sometimes compared. Though they are associated with transitivity, as the passive constructions are, the data examined here did not reveal any perspective taking functions. Their association with the inanimacy of the canonical object suggests a passive subjection to a resulting state, which is not seen in the German and English constructions.

The lack of effect for the new/given distinction was surprising. One explanation is that since there is no explicit marking of definiteness (see footnote 35 on page 82), the fronted constituents are coerced from an originally indefinite new interpretation (associated with the positions after the verb) into a definite given interpretation (associated with positions before the verb). Another is that the association with these two constructions is weak, since a more economical alternatives for object foregrounding and subject backgrounding exists (topicalisation and argument dropping – see footnotes 29 and 30 from page 80).

3.2.6 Effect of Background Knowledge on Acceptability Responses

3.2.6.1 Background

Acceptability judgements are often made in the absence of an explicit supporting context (see e.g. Wasow and Arnold, 2005, p.1485), whether these judgements are intuitions of linguists themselves, or are elicited from participants experimentally. Wasow and Arnold argue that it is not possible to judge the acceptability of a sentence without giving it an interpretation, and that the interpretation chosen may have an effect on the judgement arrived at.[126] Experimental evidence (Hagoort et al., 2004) demonstrates that inconsistencies between real world knowledge and linguistic form can have similar electrophysiological effects to syntactic incongruities within the first 500ms of online comprehension.

An extreme example among the experimental materials is the line of dialogue below, from the film *Jaws*. The correct interpretation (that the character is discussing his price, in thousands of dollars for hunting down a man-eating shark) is not evident from the isolated clause alone (3.129a).

[126]This issue is distinct from that of how discourse relations between stimulus sentence constituents and its environment, such as the newness/givenness distinction, affect sentence acceptability. For example Keller and Alexopoulou (see e.g. 2001) manipulate (constructed) textual contexts to examine their effect. In this experiment, authentic context is added or removed to investigate the effect of the quantity of textual context or depth of background understanding.

(3.129) a. I'll find him for three

b. **Quint**: ... I value my neck a lot more than three thousand bucks, chief. **I'll find him for three**, but I'll catch him, and kill him, for ten. ... For that you get the head, the tail, the whole damn thing.

Background knowledge can be explicit or implicit. For example, a Spielberg film enthusiast might recognise the scene in question above from the isolated quotation alone, or if simply prompted with the name of the film (this constitutes implicit knowledge). Alternatively the background can be supplied explicitly – a description can be supplied or the immediate textual context can be given. Alternatively, there may be interactions between implicit and explicit context: responses to a stimulus presented *with* explicit context might be additionally modulated by whether the respondent was already familiar with the context (e.g. has seen the film, and remembers the relevant scene).

In this section, the combined effect of implicit and explicit knowledge will first be investigated directly using the responses from the experiments in sections 3.2.2 to 3.2.5. As participants were asked whether they had seen the film in question, and whether they remembered the scene for each stimulus presentation, these responses can be examined for their effect on judged acceptability.

To investigate the effect of explicit knowledge, a replication of the English language experiments (sections 3.2.2 and 3.2.3) with reduced materials was performed in which the amount of textual context was varied. Further, in cases where the title was not given, participants were asked to guess what film the stimulus item came from, so that the effect of implicit knowledge could be examined independently.

3.2.6.2 Method

For this replication of the English language experiment, large numbers of respondents were not expected, and so the number of judgements received per stimulus over the full set of 87 experimental items might have been too small to reliably estimate the acceptability of any individual stimulus. A subset of 32 items were manually selected to be representative of the composition of materials as a whole, and in terms of estimated acceptability and unanimity among respondents (measured using the median and interquartile range of responses from the main experiment).[127]

The format of the experiment replication was identical with the exception of the amount of textual context supplied. For each item, participants saw one of four modes of presentation: the test sentence or clause alone; the test sentence with surrounding dialogue (usually one sentence preceding the stimulus, and one following); the test sentence with dialogue and scene description; and the test sentence with dialogue, scene description and film description (descriptions were usually between 15 and 25 words in length). Participants were again asked if they knew the film and scene question. In the cases where the film description was not given, they were also asked to guess the name of the film. The instructions are listed in appendices (page 198). Respondents were recruited from an undergraduate course in English Composition at Southeastern Louisiana University, and were asked to mark 30 sentences (again, each participant would only be presented with a single variant of original/altered pairs).[128]

[127] This subset contained the reference sentence and 5 practise items; 4 general fillers and 4 related fillers; 8 dative/benefactive test items and 10 active/passive test items. Including altered diathesis variants, there were 54 stimulus sentences.

[128] Many thanks for Dianna Laurent for her help in recruiting the participants.

Table 3.17: Effect of existing background knowledge on judgements

	English	Chinese	German
Has not seen film	-0.33 (0.5)	-0.22 (1)	-0.42 (0.6)
Has seen film but does not remember scene	-0.35 (0.4)	-0.23 (1.2)	-0.51 (1)
Remembers scene	-0.47 (0.8)	0.04 (1)	-0.42 (0.7)

Mean of normalised magnitude estimates are given, with interquartile range in brackets.

3.2.6.3 Results

Examining the interaction of implicit and explicit knowledge in the original experiments, revealed small but significant effects. Individual magnitude estimates were compared for three conditions: the participant claimed to remember the scene; only to have seen the film; and not to have seen the film. The results across three languages were mixed (table 3.17).

In the Chinese data, responses for which the film or scene was unknown were marginally lower (meaning participants marked more severely; $p < 0.001$, two-tailed Welch adjusted t-test over participant responses), and the data spread was broadly equal (meaning participants agree or disagreed to the same extent, regardless of pre-existing knowledge). Among the English data, responses for which the scene was known decreased responses fractionally, and were mixed in their relative data spread. In the German results, respondents who had seen the film, but did not remember the scene in question marked somewhat more severely, and with less consensus, than the other two conditions ($p < 0.01$).

In the modified replication experiment, the total number of sessions recorded was 38, and of these 20 were deemed valid.[129] A total of 600 judgements were returned, and the median number per stimulus sentence was 10. Three quarters of the respondents were women, and 16 of the 20 respondents reported their age to be between 16 and 25.

For some sentences, large differences in responses could be seen depending on the amount of textual context supplied. The mean magnitude estimation score given to sentence (3.129a) in isolation was -1.08 (i.e. that it was approximately twice worse than the neutral reference sentence) while with the surrounding dialogue (3.129b) the mean score was -0.42.

The amount of explicit context given did have an effect on the scores given (table 3.18). As the amount of context given was increased, the average score given also increased, and the data spread decreased, with the effect being stronger for altered sentences than for others. This indicates that participants found sentences more acceptable given the appropriate context, and that there was less difference in opinion between participants. The differences between the zero context and full context conditions was significant over all sentences ($p < 0.05$, two-tailed Welch t-test over responses, n=600), and separately for original ($p < 0.01$) and altered sentences ($p < 0.05$).

Finally, the data showed only slight effects of implicit knowledge of context (table 3.19). Here, among the judgements for which the film name was not supplied, the responses submitted by those who correctly guessed the film (125 scores), were compared to those who did not know the film, or guessed incorrectly (112). For the no-context and surrounding dialogue conditions, having the (correct) background knowledge was associated with increased scores and reduced data spread, but the effect only approached significance at $p = 0.1$ for both conditions

[129] Sessions were excluded if the respondent was not a native-speaker of English; if the altered practise items were as a whole marked as more acceptable than the authentic practise items (suggesting that the participant had inverted the scale); if the participant backtracked to change a judgement; or if the participant did not proceed beyond the initial six warm-up items.

Table 3.18: Effect of explicit context on judgements

	All (*n*=600)	**Original (*n*=341)**	**Altered (*n*=259)**
No Context (*n*=140)	-0.68 (1.17)	-0.37 (0.73)	-1.16 (2.00)
Surrounding Dialogue (*n*=146)	-0.45 (1.00)	-0.16 (0.41)	-0.83 (1.58)
***plus* Scene Description (*n*=182)**	-0.53 (1.00)	-0.25 (0.53)	-0.88 (1.38)
***plus* Film Description (*n*=132)**	-0.37 (1.00)	-0.04 (0.16)	-0.78 (1.04)

Mean of normalised magnitude estimates are given, with interquartile range in brackets.

Table 3.19: Effect of implicit context on judgements

	Correct Guess (*n*=125)	**Wrong Guess (*n*=112)**
No Context (*n*=140)	-0.23 (0.06)	-0.57 (1.10)
Surrounding Dialogue (*n*=146)	-0.41 (1.00)	-0.77 (1.20)
***plus* Scene Description (*n*=182)**	-0.67 (1.00)	-0.60 (1.10)

Mean of normalised magnitude estimates are given, with interquartile range in brackets.

3.2.6.4 Discussion

Pre-existing knowledge had subtle but significant effects on the responses given, but different patterns emerged across the three languages. There results are difficult to interpret for that reason.

Implicit background knowledge had a very slight effect on the level and spread of responses returned, only approaching significance. However, by providing the background context explicitly, the level of spread in responses could be reliably reduced, which has consequences for the statistical sensitivity of experiments.[130] If the tendencies seen in implicit context data could be confirmed in a larger study, that would also have consequences for experiment design. It would suggest that similar gains in reliability of judgements can be made by controlling the interpretation by any of these methods: using materials that respondents are sure to recognise (e.g. a piece of text mentioning *US Army* and *Baghdad*); by prompting with a title or short description; or by providing a full textual context.

3.2.7 Effect of Judgement Scales on Acceptability Responses

3.2.7.1 Background

Magnitude estimation was chosen as the judgement scale for the acceptability experiments described in order to provide cleaner data than would be obtained with categorial alternatives, such as a Likert scale. As discussed in section 3.1.2.3 magnitude estimation does not constrain participant responses to an arbitrary range, or to an arbitrary degree of granularity, and it appears to give privileged access to the internal cognitive scale that is operative during acceptability tasks (Bard et al., 1996). For this reason it is expected to provide larger effect magnitudes than a bounded scale would give (relative to data variance). Categorial methods such as Likert scales and pairwise comparison may force participants to make arbitrary responses to stimuli that fall on the margin between two categories, introducing extra variation into the data. Since many statistical tests (including the *t*-test, power calculations and confidence intervals) rely on the ratio between effect size and standard deviation,

[130] Many statistical tests of inference and power are proportional to the size of the effect and the square-root of the sample size, and inversely proportional to the size of the data spread: $\frac{d}{\sigma/\sqrt{n}}$, where d is the effect size. Thus reducing standard deviation increases the strength of statistical inferences.

maximising the former and minimising the latter should lead to stronger theoretical conclusions – that is, a smaller chance of drawing a false inference (Type I error), a smaller chance of missing a genuine effect (Type II error), and smaller margins of error (confidence intervals).

However, magnitude estimation is known to have issues of face-validity. Discussion with pilot respondents from the experiments described above (sections 3.2.2-3.2.5) revealed that they quickly got into their stride using the scale, despite finding it complex and bizarre at first sight. In a laboratory environment, in which participants are dealing directly with scientists and academics, there is considerable social pressure to perform their task well, and overcome any initial difficulties they may have. Financial or other compensation may also increase motivation. In the less formal environment of an anonymous web-administered experiment, uncompensated participants might be expected to approach the task less carefully, or simply to drop out if the task seems onerous. As the strength of some statistical tests also relies on sample size (such statistics are proportional to the square root of the sample size), it might be the case that gains in data cleanliness provided by magnitude estimation are offset somewhat by lower response rates.

In this section, two replications of the English diatheses experiment are described: one using a seven-point Likert scale, and the other pairwise comparison. These subexperiments were performed in parallel with the second tranche of recruitment (through postings to cinema-related web-forums) for the English language experiments described in sections 3.2.2 and 3.2.3. Respondents were randomly assigned to one of the judgement scale conditions (magnitude estimation, Likert scale, pairwise comparison), without being made aware that this was only one of three possibilities. Other than the rating methodology, and changes to the instructions that these entailed, the three subexperiments were identical.

The responses that each judgement scale provides can be compared in several ways. Response rates, in terms of participants who read through the instructions and started the experiment, can tell us if any of the scales have particular problems of face-validity. Completion rates of participants who do start can in turn give an indication of each scale's ease-of-use. The final number of stimulus responses received per candidate participant can then tell us what the combined effect of face-validity and ease-of-use is on the final quantities of data gathered using such scales in a web experiment.

Aggregate evaluations for each linguistic stimulus can also be compared. The goal of linguistic acceptability experiments is usually to obtain a measure for each experimental utterance that reflects the extent to which it is licensed by the common grammar of a language community (to the extent that the dialects or idiolects in question overlap). Ideally the aggregate sentence acceptability measures yielded using each judgement scale would be compared to some benchmark set of "true" acceptability values. By calculating the correlation of the set of aggregate sentence scores from each subexperiment to the benchmark, the extent to which each judgement scale approximates this "truth" can be evaluated.

In the analysis here, the magnitude estimation judgements from the first recruitment group are used as a benchmark. While not ideal, the first group results can be considered more reliable than the second tranche for two reasons: the first group of respondents were all direct or indirect acquaintances of the experimenters, and so may have approached the experiment more conscientiously; and the raw number of judgements gathered was also higher, by a factor of about 1.5, than any of the judgement scale subexperiments from the group of web-forum respondents. The groups differed in terms of dialect, with British Isles speakers of English predominating in the first, and North American speakers in the second, but this difference applied equally to all three judgement scale

subexperiments.[131] However, the magnitude estimation subexperiment does have an advantage over the other two methods, since it shares the scale's subtle biases with the benchmark (as discussed on page 67 of section 3.1.2.3).

Finally, as noted above, the motivation for maximising effect size, minimising data spread, and maximising sample size, is to increase statistical sensitivity. Accordingly, all three data sets were evaluated in terms of how strongly they confirmed theoretical hypotheses. Five hypotheses were chosen that are both predicted by their relevant literature, and which received support from the English language experiments in sections 3.2.2 and 3.2.3. These hypotheses were (roughly in order of increasing subtlety): that the set of original sentences are more acceptable than the altered diathesis variants; that the dative diathesis is more successful if its indirect object is animate, and it is given in the discourse; and that the passive diathesis is more successful if its subject (the direct object in the canonical active variant) is animate, and its *by*-object (the canonical subject) is new to the discourse.[132]

3.2.7.2 Method

As already discussed in section 3.2.2, the second group of participants was recruited via postings to over 20 British and North American online forums and discussion groups that dealt with cinema. Respondents who followed a link were randomly assigned to one of the three judgement scale conditions, and were asked to mark 15 items, presented one per page. In all cases the order and composition of experimental items were balanced and randomised, as before, and all sessions started with the same five practise sentences.

For the magnitude estimation subexperiment, the same reference sentence from Woody Allen's film *Annie Hall* was presented together with each subsequent item, and participants were asked to assign a score on the basis of their perceived ratio judgement of quality relative to that reference.

In the Likert scale subexperiment, sentences were marked on an absolute 7-point numerical scale.[133] Explicit descriptions were not supplied for each point on the scale, but examples were given in the instructions that characterised 7 as a perfect sentence, 5 as an unwieldy sentence, 3-4 for an illformed sentence and 1-2 for an uninterpretable sentence.

In the pairwise subexperiment each sentence was evaluated with respect to the item that preceded it (*not* relative to a single reference item as is the case in magnitude estimation). Participants could indicate that they found the current item "better", "the same" or "worse" than the preceding item. These terms were characterised again in terms of awkwardness, illformedness and difficulty of understanding.

Instructions were kept as uniform as possible. Due to the numerical manipulations involved in magnitude estimation, the set of instructions were marginally longer than the other two. All three sets of instructions are included in appendix B (pages 192 to 196).

[131] The difference in proportions of British Isles dialects to North American dialects between the first recruitment group, and the three subexperiment groups was significant ($\chi^2 = 29.2,\ df = 3,\ p < 0.001$). There was no significant difference in this proportion among the three subexperiment groups ($\chi^2 = 1.38,\ df = 2,\ p < 0.5$).

[132] The effect of the last three hypotheses only approached significance at $p < 0.1$.

[133] The median number of gradations used by each participant in the original experiment was 8 and the median ratio between the lowest and highest score assigned by a single participant was 7.5 (though some participants expressed differences in acceptability of 100-fold or more). Accordingly, a choice of 7 Likert gradations might be expected to constrain the responses of some participants (some have advocated figures as high as 20 as the ideal number of ordinal categories; see Schütze, 1996, p.78).

Table 3.20: Comparison of response rates to Likert, Magnitude Estimation and Pairwise Comparison measurement scales.

Responses	Likert	Mag. Est.	Pairwise
Number of candidates assigned	129	153	140
Number of participants to complete pre-survey	74	86	78
Number of participants to start experiment	65	50	65
Number of participants to submit valid sessions	49	41	51
Number of judgements per valid session	21	21	19
Median Time per Item (seconds)	19	20	20
Number of judgements returned	1039	871	974
Number of judgements per candidate	8.1	5.7	7.0

3.2.7.3 Results

Response rates and time spent As the results in table 3.20 indicate, fewer candidates assigned to the magnitude estimation subexperiment chose to participate, relative to the other two judgement scale conditions. In all three conditions between 50 and 60% of candidates who were directed to the welcome page went on to complete the pre-survey. Of those about 85% continued through the instructions and commenced the experiment (i.e. submitted at least one score) for the Likert and pairwise variants; for magnitude estimation less than 60% continued (the effect was not significant).

The next two lines of table 3.20 describe the valid sessions for each subexperiment: as before, a session was deemed invalid if the participant appeared to have inverted the marking scheme (by marking original items worse as a whole than altered variants among the five initial warm-up items), if the participant backtracked during the experiment to change a score, or if the participant failed to progress beyond the warm-up items.[134]

Both the number of judgements per valid session and the time spent per response are almost equal, suggesting that all three judgement scales were similarly easy to use.

However the final volume of data received per candidate assigned to the magnitude estimation variant of the experiment was 30% less than that received using the Likert scale ($\chi^2 = 7.4,\ p < 0.01$), presumably due to face-validity and/or the increased length of the instructions. The difference in data received from the pairwise experiment was not significant.

Correlation to Benchmark Sentence Scores For the analysis of responses, some additional sessions were excluded: those that were submitted by non-natives of the language, and sessions that were judged to contain mistaken or facetious responses.[135] The set of response data from the first recruitment group (acquaintances recruited via email, before the Likert and pairwise comparison subexperiments were added) amounted to 1473 individual judgements in 48 valid user sessions. The five practise sentences were excluded from this analysis, leaving 1233 judgements over 148 sentences, and giving a median of 8 judgements per sentence variant.[136] For each of these 148 sen-

[134]These response rate figures include some sessions that were excluded from the analysis of responses: those of non-natives, and sessions that appeared to contain outlier responses that were judged to be mistaken, or facetious. The mode age range reported by respondents was 21-25 years. The lowest age range reported was 16-20, and the highest 76-80. The mean age of respondents was estimated at 30 years (by taking the middle value from each range). Female respondents were outnumbered by males, by a factor of four.

[135]Two sessions were excluded after being identified as containing implausibly extreme values: one from the benchmark set, and one from the magnitude estimation. They have been discussed in on page 96 of section 3.2.2.4. No such sessions were found among the pairwise comparison and Likert scale experiments, as these rating methods constrain participant responses.

[136]The practise sentences are unrepresentative for several reasons. Firstly, they were selected for every user session, and so received much larger numbers of individual judgements. Secondly, the initial judgements in any session can be expected to

tences, an aggregate score was calculated. The literature on magnitude estimation suggests that the mean is the appropriate measure for cross-participant aggregation (see section 3.1.2.3), but because of the anormality of the data returned, the median was taken also.[137] Aggregate scores for the magnitude estimation scale subexperiment were calculated identically, on the basis of 624 individual (log-normalised) scores for non-practise items (median 4 judgements per sentence variant).

Aggregate scores for the Likert data were calculated using the mean, over 739 non-practise responses (median 5 judgements per item).

Pairwise aggregate scores were calculated as a their percentile acceptability over 689 non-practise responses (median 4 per item). Individual pairwise scores of "better" (than the preceding random reference) were given the value 1, "same" was counted as 0.5, and "worse" counted as zero. For a single sentence, the mean of these values gives its percentile acceptability, relative to the set of 148 non-practise items.

A series of correlation coefficients were calculated. Pearson (linear) and Spearman (rank) were both used, since the interval status of aggregate Likert data remains controversial (linear correlations assume interval data). Since some of the judgement methods may be sensitive to more or less subtle effects of acceptability, each data set was split into three acceptability terciles, and correlations were additionally calculated for these (the set of 148 sentences were ordered by their benchmark scores, and each was allocated to the upper, medium or lower acceptability tercile).[138]

The first three lines of table 3.21 compare the correlation over all sentences between the benchmark data, and each of the three sets of judgement scale data, using the mean to compute individual sentence scores.[139] It suggests that the Likert scale data approximates most closely to the benchmark data, with a correlation coefficient of 0.74, though the difference between this figure and that for magnitude estimation is not significant. The pairwise comparison data does correlate significantly less than both the magnitude estimation and Likert data ($p < 0.05$ and $p < 0.01$ respectively, two-tailed t-tests, n=148).[140] Within individual acceptability terciles, none of the differences between judgement scale conditions were significant. Using the median to calculate sentence scores lowered correlations, but the relative ordering between judgements scales was preserved.

The first three lines of table 3.22 show the same correspondences computed with the rank (Spearman) correlation. This reduces the differentials between the three techniques.

Though the magnitude estimation data had been examined to exclude mistaken or facetious responses, its performance relative to the Likert scale data may have been compromised by some genuine, but extreme, individual responses. The fact that the rank correlation figure is higher than the linear correlation suggests that this may be the case (though the difference is not significant). To test this, an arbitrary cutoff was applied the magnitude estimation responses. All responses that lay three standard deviations above or below the mean, were discarded.[141] Though this only amounted

be more noisy, as the participant is still settling into their task. Finally, for the pairwise comparison data, the percentile scores calculated for practise sentences are not comparable to those for others – since the method of presentation ensured that the practise items always appeared as a block at the start of the session, their percentile scores express acceptability *relative to the set of practise items*, rather than relative to the materials as a whole.

[137] See figure C.1 on page 217 of the appendices for a histogram demonstrating the anormality of the data returned with the magnitude estimation technique.

[138] The same correlations were also calculated using the median as aggregate measure for the Likert and pairwise data, yielding very similar patterns.

[139] Significance levels indicated as follows: $*$, $p < 0.05$; $**$, $p < 0.01$; $***$, $p < 0.001$.

[140] Significance of difference for two dependent correlations, calculated with t-test, two-tailed:
$t = (r_{xb} - r_{yb}) \sqrt{\frac{(n-3)(1+r_{xy})}{2(1-r_{xb}^2-r_{yb}^2-r_{xy}^2+2r_{xb}r_{yb}r_{xy})}}$, where n is the number of data points being compared, r_{xy} is the correlation between two measurement scales, and r_{xb}, r_{yb} are their correlations with the benchmark.

[141] The mean log normalised score for the magnitude estimation subexpriment data was -0.5, and the range of plus/minus three standard deviations spanned -3.6 to $+2.6$. So that the correlation yielded would be directly comparable to the existing

Table 3.21: Pearson linear correlations between benchmark data and three judgement scales

Pearson correlation, *r*, against Benchmark	**All**	**Upper**	**Middle**	**Lower**
Magnitude Estimation	0.60^{***}	0.07	0.33^{*}	0.55^{***}
Likert	0.74^{***}	0.20	0.46^{***}	0.63^{***}
Pairwise	0.61^{***}	0.32^{*}	0.40^{**}	0.58^{***}
Magnitude Estimation, Trimmed	0.70^{***}	0.32^{*}	0.40^{**}	0.53^{***}

Mean is used as measure of central tendency for magnitude estimations.
Upper, middle and lower acceptability terciles are compared.

Table 3.22: Spearman rank correlations between benchmark data and three judgement scales

Spearman correlation, ρ, against Benchmark	**All**	**Upper**	**Middle**	**Lower**
Magnitude Estimation	0.65^{***}	0.25	0.35^{*}	0.62^{***}
Likert	0.67^{***}	0.18	0.42^{**}	0.66^{***}
Pairwise	0.59^{***}	0.24	0.29^{*}	0.59^{***}
Magnitude Estimation, Trimmed	0.72^{***}	0.36^{*}	0.40^{**}	0.61^{***}

Mean is used as measure of central tendency for magnitude estimations.
Upper, middle and lower acceptability terciles are compared.

to 10 responses of more than 600, there was a clear increase in correlations. The linear correlation (labelled "Magnitude Estimation, Trimmed") rose to just below the level for the Likert data, and the rank correlation climbed above that for the Likert data.

To examine the effect of response rate (and so face-validity), the linear correlation between the benchmark and each judgement scale data set was also calculated incrementally, in the order that responses were submitted by participants. The development of these correlations are plotted in figure 3.3. All three start at full correlation (any correlation over two points will result in a correlation coefficient of $+1$ or -1), and start to stabilise around 300 judgements (or 2 judgements per sentence variant). The Likert data gives a consistently higher correlation than the pairwise comparison or full magnitude estimation data. While all three data sets are affected by outlier responses (indicated by abrupt jumps in the correlation), the magnitude estimation subset seems to be most vulnerable to fluctuations. Trimming out the small number of responses that fell outside three standard deviations of the mean within the magnitude estimation data improved correlations. It is not clear whether any of the three methods have a clear upward trend (which would suggesting continuing convergence towards the benchmark). Individual plots for each of the three acceptability terciles are supplied in the appendices (see pages 220 to 221), though as noted, most of the differences within these acceptability ranges are not significant.

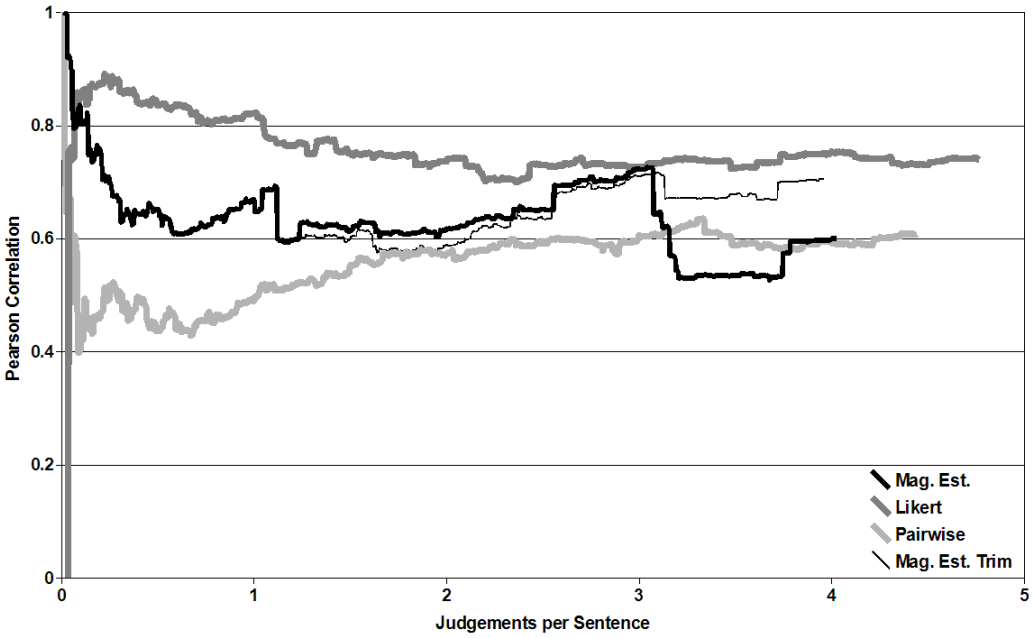
Confirmation of Experimental Hypotheses Five hypotheses were tested over the experimental results for the benchmark data, and each of the three judgement scales. For each hypothesis, experimental items were grouped by a binary factor to determine if it had a significant effect on aggregate acceptability scores.

Group means of aggregate acceptability scores for 85 original (authentic) sentence items were compared to those for 70 altered (i.e. diathesis variant) sentence items.[142] The 24 dative/benefactive diathesis items were grouped by two factors: the animacy of the indirect object, and its information

figures, the threshold was not applied to the benchmark data.

[142]Practise items were not excluded from this analysis. Two of the 70 altered sentences were not diathesis variants, but rather practice items that had been manually altered to be intentionally unacceptable.

Figure 3.3: Correlation of data from three measurement scales to a magnitude estimation benchmark



Pearson (linear) correlation over acceptability stimuli between data gathered using three rating scales.
Trimmed magnitude estimation data set contains only responses within range of plus/minus three standard deviations.

status in the discourse (new or given). The 30 active/passive diathesis items were grouped by the animacy of the canonical direct object (which appears as the subject in the passive variant) and and givenness of the canonical subject (the *by*-object in the passive). For each of the dative/benefactive, and active/passive items, the success of the diathesis was estimated by taking the difference of the mean aggregate scores for original and altered variants.[143] Two of the five hypotheses had been previously confirmed by the English acceptability experiment at conventional levels (the original/altered effect, and the effect of animacy on the dative/benefactive – see section 3.2.2), using all magnitude estimation responses (i.e. both the first recruitment group of acquaintances, and the second group of web-forum respondents). The other three hypotheses only approached significance (at $p < 0.1$; see sections 3.2.2 and 3.2.3.4).

Significance levels were calculated by taking a two-tailed independent samples *t*-test, with unpooled variance, and a Welch correction to degrees of freedom for unequal variances.[144] The non-parametric Mann-Whitney/Wilcoxon test of independent samples was also taken.

Table 3.23 shows the degree to which each hypothesis was confirmed (or otherwise) by each data set. Since some of the *p*-values would be vanishingly small, the Welch test *t*-scores, and the Mann-Whitney/Wilcoxon test *z*-scores are given directly, together with an indication of conventional significance levels ($*$, $p < 0.05$; $**$, $p < 0.01$; $***$, $p < 0.001$.). For the relatively coarse difference between original and altered sentences, and the large effect of animacy on the dative/benefactive items, the pairwise comparison data confirmed hypothesis to the highest degree of confidence, with or without parametric assumptions.

None of the three subexperiment sets detected the effect of animacy on the productivity of the

[143] Similar results to those described here were obtained by making the same comparisons using the median to compute aggregate scores.

[144] Welch (1938): $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, $df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$.

Table 3.23: Comparing means across groups of sentences

	Benchmark Magnitude Estimation	**Replicated Magnitude Estimation**	**Likert**	**Pairwise Comparison**	**Trimmed Magnitude Estimation**
Original (*n*=85) vs Altered (*n*=70)	$6.4^{***}(6.5^{***})$	$5.9^{***}(5.7^{***})$	$6.6^{***}(5.8^{***})$	$8.4^{***}(7.0^{***})$	$7.0^{***}(6.2^{***})$
Dative animate (*n*=12) vs inanimate (*n*=12) indirect object	$3.2^{**}(2.6^{**})$	$2.4^{*}(2.7^{**})$	$3.3^{**}(2.6^{**})$	$4.6^{***}(3.4^{***})$	$4.0^{***}(3.3^{***})$
Passive animate (*n*=15) vs inanimate (*n*=15) subject	1.6(1.5)	0.3(0.5)	0.6(0.8)	0.2(0.2)	0.8(1.2)
Dative given (*n*=10) vs new (*n*=14) indirect object	$2.9^{**}(2.0^{*})$	1.1(1.4)	$2.2^{*}(2.0^{*})$	1.5(1.5)	$1.9^{*}(1.8^{*})$
Passive given (*n*=22) vs new (*n*=8) *by*-object	$3.6^{***}(2.9^{**})$	$2.3^{*}(2.5^{**})$	$2.5^{*}(2.3^{*})$	1.6(1.4)	$3.0^{**}(3.0^{**})$

Main figures are t-scores from a Welch corrected independent samples t-test over sentence scores and difference scores. Figures in brackets are z-scores from a Mann-Whitney/Wilcoxon test of independent samples.

passive. The Likert data was most sensitive to the effect of givenness on the dative/benefactive diathesis, and magnitude estimation was most sensitive to its effect on the passive, once outlier responses had been discarded.

The same statistics can be calculated over individual responses from each data set, rather than using aggregate scores for each sentence item. This results in much larger sample sizes, and so can be expected to increase significance levels, though this may be offset somewhat by the increased variance among individual responses (the taking of aggregate measures, as in the analysis above, eliminates the variance among responses to a single item). Since no single participant was presented with both diathesis variants of a single experimental item, individual responses for the difference in acceptability between original and altered variants are not available. For this reason, the simple scores for altered variants were used to evaluate the success of diathesis items.

Table 3.24 shows the significance levels for tests of the five hypotheses over individual responses (at these samples sizes, the t-scores are identical to z-scores, so the parametric and non-parametric results are directly comparable). The p-values produced for some of the non-parametric tests were beyond the precision of the software package used, and returned z-values of infinity. Here the Likert scale data provided strongest confirmation on the original/altered distinction and the effect of givenness on the dative/benefactive diathesis and the pairwise data was strongest on the effect of animacy for the dative/benefactive. The trimmed magnitude estimation data best detected the effect of animacy on the passive (that had not been confirmed by the sentence-level analysis), and the untrimmed data was strongest on the effect of information status on the passive.

To eliminate the effect of differences in samples size on statistical sensitivity, and so factor out the effect face-validity, the z-scores just seen can be adjusted to a notional uniform sample size of 1000 responses.[145] With this adjustment, either the Likert or pairwise data yield a higher significance value on three of five hypotheses.

[145] Assuming the same estimates of effect size and standard deviation, each z-score is multiplied by a factor of $\sqrt{1000}/\sqrt{n}$, where n is the actual sample size.

Table 3.24: Comparing means across groups of judgements

	Benchmark Magnitude Estimation (*n*=1473)	**Replicated Magnitude Estimation (*n*=809)**	**Likert (*n*=979)**	**Pairwise Comparison (*n*=929)**	**Trimmed Magnitude Estimation (*n*=794)**
Original (55%) vs Altered (45%)	$15.2^{***}(\infty^{***})$	$13.6^{***}(\infty^{***})$	$18.2^{***}(\infty^{***})$	$16.8^{***}(\infty^{***})$	$14.7^{***}(\infty^{***})$
Dative animate (50%) vs inanimate (50%) indirect object	$5.7^{***}(4.4^{***})$	$1.8^{*}(1.6)$	$4.7^{***}(3.9^{***})$	$4.7^{***}(4.5^{***})$	$2.2^{*}(1.6)$
Passive animate (50%) vs inanimate (50%) subject	$4.8^{***}(5.5^{***})$	$3.4^{***}(3.8^{***})$	$1.9^{*}(1.6)$	$2.9^{**}(2.9^{**})$	$4.0^{***}(3.9^{***})$
Dative new (58%) vs given (42%) indirect object	$4.0^{***}(2.0^{*})$	$1.0(1.1)$	$4.1^{***}(3.1^{**})$	$3.8^{***}(3.7^{***})$	$1.3(1.1)$
Passive given (73%) vs new (27%) subject	$3.5^{***}(2.8^{**})$	$3.7^{***}(3.0^{**})$	$1.0(0.6)$	$1.8^{*}(1.8^{*})$	$3.2^{***}(2.8^{**})$

Main figures are z-scores from a Welch corrected independent samples t-test over participant responses.
Figures in brackets are z-scores from a Mann-Whitney/Wilcoxon test of independent samples.

Table 3.25: Comparing means across groups of judgements, normalised for sample size

	Benchmark Magnitude Estimation (*n*=1473)	**Replicated Magnitude Estimation (*n*=809)**	**Likert (*n*=979)**	**Pairwise Comparison (*n*=929)**	**Trimmed Magnitude Estimation (*n*=794)**
Original (55%) vs Altered (45%)	$12.5^{***}(\infty^{***})$	$15.1^{***}(\infty^{***})$	$18.4^{***}(\infty^{***})$	$17.4^{***}(\infty^{***})$	$16.5^{***}(\infty^{***})$
Dative animate (50%) vs inanimate (50%) indirect object	$4.7^{***}(3.6^{***})$	$2.0^{*}(1.8^{*})$	$4.8^{***}(3.9^{***})$	$4.8^{***}(4.7^{***})$	$2.5^{*}(1.8^{*})$
Passive animate (50%) vs inanimate (50%) subject	$4.0^{***}(4.5^{***})$	$3.8^{***}(4.2^{***})$	$1.9^{*}(1.6)$	$3.0^{**}(3.1^{**})$	$4.5^{***}(4.3^{***})$
Dative new (58%) vs given (42%) indirect object	$3.3^{***}(1.7^{*})$	$1.1(1.2)$	$4.1^{***}(3.1^{***})$	$4.0^{***}(3.8^{***})$	$1.5(1.2)$
Passive given (73%) vs new (27%) subject	$2.9^{***}(2.3^{*})$	$4.1^{***}(3.3^{***})$	$1.0(0.6)$	$1.8^{*}(1.8^{*})$	$3.6^{***}(3.1^{***})$

Main figures are z-scores from a Welch corrected independent samples t-test over participant responses.
Figures in brackets are z-scores from a Mann-Whitney/Wilcoxon test of independent samples.

3.2.7.4 Discussion

As expected, the magnitude estimation rating subexperiment had a lower response rate that the other two methods, presumably due to its relative lack of face-validity. The number of judgements returned per candidate participant was significantly lower than with the Likert scale.

Using a range of correlation metrics, the data gathered using a Likert scale approximated the benchmark set of acceptability scores considerably more closely than did either magnitude estimation or pairwise comparison. This difference in convergence towards the benchmark is not accounted for by the varying quantities of data gathered with each technique. However, when a small number of apparently genuine but extreme estimates were discarded, the magnitude estimation data approached the performance of the Likert scale data on a linear correlation, and exceeded it on a rank correlation.

Of five sample experimental hypotheses that had been received support from the benchmark data, and which represented a range of effect sizes, both the pairwise comparison and Likert data provided equal or stronger support for three of the hypotheses than that provided by the magnitude estimation data, regardless of any parametric assumptions. This could not be accounted for by differences in response rates.

These results are surprising. Bard et al. (1996) demonstrate that magnitude estimates can access perceived acceptability with a high degree of accuracy. The methodology used here followed recommendations from the literature (Stevens, 1975; Lodge, 1981; Bard et al., 1996). Though magnitude estimation should yield interval data, some of the results seem more consistent with an ordinal scale, or with an interval scale whose distribution is distorted by outliers. Ordinarily, a limited scale (such as Likert or pairwise comparison) would be expected to compress real differences, and perform worse in linear correlations, relative to an open-ended scale. But in these experiments magnitude estimation compared more favourably to the other two judgement methods on rank correlations. On some of the hypothesis tests, results were more significant with nonparametric than parametric tests.

The experiments presented here had limitations. Though numbers of participants were not small, the numbers of judgements per stimulus sentence (four or five) were lower than would be ideal – the mean of magnitude estimates for a single experimental item needs between 10 and 20 judgements to stabilise (see figure 3.1 on page 96). The predominant dialects within the first and second recruitment differed, which may account for the relatively modest correlations to the benchmark (though this limitation applied equally to all three subexperiments) and the fact that there is not a clear trend towards full convergence. And the experimental materials were biased towards particular constructions, in a single language.

However, if these results can be replicated, they would have implications for experimental design. They would indicate, not only that the face-validity of magnitude estimation leads to lower response rates in web-based experiments, but also that the data gathered in no less noisy than can be obtained with a simple Likert scale. Though this is only indicated for the relatively small numbers of judgements per item seen here, it is precisely under such conditions that the sensitivity of magnitude estimation should come to the fore. And this performance can only be achieved by discarding participant responses by some arbitrary threshold (such as the range of plus/minus three standard deviations used here).

While magnitude estimation is widely used and well established methodology for acceptability testing in linguistics (see e.g. Brennan et al., 1975; Shapiro, 1997; Featherston, 2005; Russell et al., 2005; Myers, 2007) I am not aware of any published study that directly compares its efficacy against

other judgement scales on grammatical phenomena.[146] A quick review of such studies in other fields revealed work that found category scales to yield data that is equally informative (e.g. on judgements of occupational prestige, Orth and Wegener, 2006; and ratings of food sounds on a scale of pleasantness, Vickers, 1983) or superior (e.g. subjective ratings of well-being, Kaplan et al., 1979; and rating the flavour of foods, Lawless, 1989) to that yielded by magnitude estimation scales on non-physical phenomena. A survey by Wegener (1983) over a range of sensory and social judgement tasks also found that both category and magnitude estimate ratings yield interval scales. Only one study found magnitude estimation to be more sensitive (see Lodge, 1981, p.79). This is then, a question that deserves more attention.

To try to discover what accounts for the difference between the performance seen here and the claims made in the literature for magnitude estimates of linguistic acceptability, several experiments could be carried out. The experiment should be replicated with larger numbers of participants, and perhaps other materials, to determine if the results are simply due to unusually biased groups of respondents, or to an abnormal set of stimuli. The experiment could also be replicated in a laboratory to determine if the choice of web-delivery is to blame, or existing lab experiments could be replicated over the web. In particular, participants that submit extreme responses could be interviewed, to determine whether these were mistaken, or accurately reflected individual differences in perceived acceptability.

The choice of response modality may also play a role. In these experiments, levels of education and numeracy were not controlled for. A replication with participants who varied in level of education and numeracy could determine what, if any, noise is being introduced to the data by difficulties in manipulating numbers.[147] And some pilot participants reported that they developed ad-hoc acceptability categories over the course of the experiment for the items they were asked to rate. In the full experiment, participants generally only used about 7 different gradations over a set of up to 30 experimental stimuli, suggesting that they were approaching the task in a somewhat more systematic and less impressionistic way than was hoped. A numerical response scale facilitates the use of such ad-hoc categories, since particular points on a number scale are memorable in a way that points on other modality scales such as loudness or line length are not. Finally, an experiment that introduced one or more calibration or training sessions using other stimulus types (e.g. physical stimuli such as brightness or loudness; or attitudinal tasks from the social sciences literature) could investigate what role illustration and task familiarity play.

3.2.8 General Discussion

In the acceptability experiments on diatheses in the three languages, some earlier theoretical arguments from the first two chapters were confirmed. It is clear that such constructions are not exclusively licensed at the lexical level. For example, accounts that claim that receipt and possession are necessary for the dative double object construction (e.g. Pinker, 1989), are challenged by successful sentences such as *bring you your son back*. If such abstract receipt is acceptable for the double object construction, it is unclear where the boundary of the formal category of possession should be drawn. Equally, similar or identical senses of a single verb proved to be more or less successful, as

[146] Southwood and Flege (1999) do such a comparison on judgements of accentedness, finding similar results for category and magnitude estimation methods. Antonella Sorace's (1992) thesis contains a comparison based on syntactic phenomena (personal communication), but this document was not available to the author.

[147] It seems unlikely that the predominantly young, male, computer-literate respondents to this study have particular problems with numeracy, relative to the population at large.

did different verbs that have very similar semantics. Further, the effect of semantic typing information that could be plausibly stored at the lexical level (cf. selectional restrictions) such as animacy, appeared to be mediated by interpretation. For example the admissibility of inanimate nouns that can denote groups of people as indirect objects in the double object construction varied, suggesting that plausibility played a role – the people at *the main office* are plausible recipients of *a wire*; the people of *Casablanca* are not such plausible recipients of Humphrey Bogart's character *Rick Blain* (see 3.2.2.5). Further, section 3.2.6 demonstrated that the perceived acceptability of sentences is modulated both by explicit textual context and by the reader's state of knowledge.

A range of semantic and discourse features, both of verbs and of their arguments, were shown to have significant effects on the acceptability (and presumably productivity) of diatheses in the three languages. Transitivity (including telicity, causation by the subject and affectedness in the object) had an effect on the German passive, and on the Chinese *ba* and *bei* constructions. The given/new distinction had an effect in the German passive, the English passive and the English dative/benefactive diatheses. Negative affect, associated with the event as a whole, or particular to its consequences for the canonical object, was found to be associated with the Chinese *bei* construction, as expected, but also with the *ba* construction and the German passive construction. Benefit (positive affect) operated on the canonical subject of the German passive and the indirect object of the English dative/benefactive. Animate and/or aware participants were preferred for earlier sentence positions in the German and English diatheses experiments, which is consistent with ideas of an anthropocentric saliency hierarchy.

Crucially, the effect on the relative acceptability of diathesis variants of such semantic and discourse features was sometimes of the same order of magnitude as clear syntactic errors, such as those in the practise sentences (reproduced below, with median magnitude estimates in brackets). Among the 100 Chinese altered *ba* and *bei* items 39 received median scores of -1.00 or less, comparable to the scores for the syntactically illformed (3.130) and (3.130); among the 50 German altered passive items, 15 were given scores of -1.32 or less, comparable to (3.132) and (3.133) (in the English materials only one sentence was rated as similarly illformed to either of the clear syntactic violations).

(3.130) gei wo ba touxiang (-1.00)
give me grip idol
'Throw the idol me'
给我把头像

(3.131) ni you bu yao xiang dao xing le (-1.00)
you again not will think at sex PFV
'Stop thinking sex about'
你又不要想到性了

(3.132) Wirf das Götzenbild mir (-2.00)
throw the idol to-me
'Throw the idol me '

(3.133) Hör endlich immer nur an Sex zu denken auf (-1.32)
give at-last always only on sex to think up
'Stop thinking on sex about '

In summary, a wide range of semantic and discourse features operate in the productivity of common diatheses. There is considerable overlap among these three languages in the set of factors that have

an effect. And the effect of structural violations does not appear to be privileged over those of semantic and discourse properties.

3.3 Corpus Validation Exercise

3.3.1 Background

Judgement experiments are necessarily limited in scale, due to the effort involved in preparing suitable materials by hand, and in recruiting participants. However, such small quantities of materials can be hand-coded for very subtle semantic and discourse distinctions – for example, the difference between sentience (i.e. consciousness) and awareness (of a particular event) can be judged.

Automated corpus analysis is capable of processing much larger quantities of text, but the features that are directly recoverable are limited to structural and lexical properties. The features of interest to this study may not be directly recoverable (those emerged from the experiments in sections 3.2.2 to 3.2.5: volition and awareness; cause and effect; telicity and duration; new and given information; positive and negative affect; anthropocentric saliency). For example, to detect volition, one might try to search for adverbs such as *deliberately*, and ascribe volition to the subject of all such sentences. However, such adverbs often seem redundant with volitional verbs (consider *?she deliberately asked directions*) and are more likely to appear with what may be typically non-volitional verbs (*he deliberately crashed the car*).

In this corpus exercise a subset of the features of interest to us are searched for using structural correlates. While these detectable textual features might be imperfectly associated with the features of interest, interpretable patterns may emerge over such large quantities of text. In this way, the corpus data may provide rather coarse, but more robust (statistical) support to some of the conclusions from the judgement experiments.

A random 15 million word sample of the British National Corpus (BNC) is examined here. One of the questions to be investigated is whether the features of interest pattern most strongly with surface grammatical role, with canonical grammatical role or with serial ordering of arguments. The Minipar lexicalised dependency parser (Lin Dekang, 1998) is used here, as it incorporates a named-entity recogniser, and achieves a high level of accuracy when identifying the arguments of a verb – both in surface and canonical terms. It is also better suited to the BNC, as it has been developed for general text, rather than for a restricted domain (most statistical parsers such as those of Collins or Charniak are trained exclusively on text from one financial newspaper, the *Wall Street Journal*).

The structural features that were examined included person (i.e. 1st, 2nd or 3rd person), animacy and concreteness, all of which may correspond to anthropocentric saliency. The first is directly recoverable from text, and the other two can be estimated from WordNet lexicon entry (Miller et al., 1990) for the head nouns of verbal arguments. Each of these properties are expected to be associated with the earlier argument positions of the sentence.

Animacy is also associated with particular kinds of participation (volition, causation), and so is expected to prefer the positions of grammatical and canonical subject positions to others.[148]

Definiteness of a nominal phrase will be viewed as an indicator of its being given in the discourse (Biber et al., 1999, pp.237,258f), so again, such phrases are expected to be more strongly associated with the head of the sentence.

[148] "Grammatical" function positions refer to surface function, and "canonical" positions refer to "deep" or "logical" positions.

Table 3.26: BNC composition

	Texts	Sentence Tokens	Word Tokens
Written	3209	5188373	89740544
Spoken – Content-Governed	762	430348	6154248
Spoken – Demographic	153	612049	4211216
Total	4124	6230770	100106008

The length of arguments is taken as a measure of constituent heaviness, which is associated with the end of the clause.

Positive or negative affect (as recorded in the Whissell dictionary of affect norms) is examined as it has been suggested to play a role in passive-like diatheses in several languages.

Finally the proportion of progressive verb forms (which is associated with atelicity) is tested for association with the passive and active forms.

3.3.2 Materials

The corpus used for the this study was the British National Corpus version 1.0 (Burnard, 1995)[149], a 100 million word balanced corpus of modern English from the United Kingdom, comprising over 4,000 texts of up to 45,000 words each.[150] The entire corpus is tagged using the CLAWS5 tagset (Leech et al., 1994), and stored in SGML markup form (Corpus Interchange Format; Burnage and Dunlop, 1992). A syntactic parser is being used in this exercise, so all tagging and markup was discarded before processing.

The written component of 90 million words is balanced for domain between imaginative and informative texts (ca. 25% and 75% respectively), and for medium between books (60%), periodicals (25%), and other written materials (15%). The texts date from between 1975 and 1994, except for a small number of popular imaginative works going back to 1964. Texts were also chosen to be representative in terms of topic area, text length, age and gender of target audience, level of difficulty and personal attributes of the author, such as age, gender and region.

The spoken component (ca. 10 million words) is split into two parts. The demographic part consists of recorded day-to-day conversations from 124 speakers from around the UK, balanced for gender, age and socio-economic status. The context-governed part contains more formal types of spoken language, balanced across four categories: informational and educational events; business events; institutional and public events; and leisure events. An overview of the contents is given in the table 3.26.

For reasons of computing capacity, it was not possible to process the entire corpus in this exercise. A random subset of the texts was selected, that amounted to 15% of the total word-count of the corpus.

3.3.3 Procedure

3.3.3.1 Preprocessing

The raw corpus text was parsed with Minipar. Its named entity recognition function recognises people and groups (e.g. corporations, governmental organisations), locational and temporal categories

[149]Copyright notice: "Data cited herein has been extracted from the British National Corpus Online service, managed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved."

[150]Full listing available at http://www.natcorp.ox.ac.uk/docs/userManual/bncIndex.html [viewed 30th July 2006].

of various sorts (country, city, time, date), and other quantities such as prices, amounts, phone numbers and addresses. The Minipar entity word lists were augmented with the more comprehensive lists supplied with the GATE natural language processing application framework (Cunningham et al., 2002).[151]

An evaluation of Minipar (Lin Dekang, 1998) claims high accuracy for the detection of subjects (precision of 89%, recall of 79%) and objects (precision of 88%, recall of 72%), using the Susanne hand-parsed corpus (Sampson, 1995) as a benchmark. Its performance on prepositional phrase attachment is also respectable (precision of 78%, recall of 72%).[152]

In this exercise Minipar parsed a variety of structures successfully, including *wh*-movement, topicalisation, control and raising, double object constructions and the passive. Minipar annotates verbs for voice, tense and morphological aspect (the perfective and continuous forms). Arguments are annotated for both grammatical (surface) and canonical (logical) structure. Grammatical (i.e. surface) subjects are marked with the dependency relation *s,* as is seen to hold between the token *He* and the verb *seem* in node 1 of the dependency parse in (3.134).[153] The gapped subject governed by the verb *do* is represented by node E2, where *he* is the canonical subject (i.e. logical subject, dependency relation *subj*), and this is linked back to its antecedent in node 1.

(3.134) He seemed to do it

```
(
E0 (<emptyToken> fin C <emptyParent>)
1 (He he N 2 s (gov seem))
2 (seemed seem V E0 i (gov fin))
3 (to to Aux 4 aux (gov do))
4 (do do V 2 sc (gov seem))
E2 (<emptyToken> he N 4 subj (gov do) (antecedent 1))
5 (it it N 4 obj (gov do))
)
```

The parsed structure of a passive works along similar lines (3.135). *It* is both the grammatical subject (*s*) and the canonical object (*obj*) of the verb *give*, while *me* is a prepositional object (*p-comp*) and canonical subject (*by-subj* for passives).

(3.135) It was given to him by me

```
(
E0 (<emptyToken> fin C <emptyParent>)
1 (It it N 3 s (gov give))
2 (was be be 3 be (gov give))
3 (given give V E0 i (gov fin))
E2 (<emptyToken> it N 3 obj (gov give) (antecedent 1))
4 (to to Prep 3 mod (gov give))
5 (him him N 4 pcomp-n (gov to))
6 (by by Prep 3 by-subj (gov give))
7 (me me N 6 pcomp-n (gov by))
)
```

[151] GATE is a modular application that can accommodate Minipar as a unit, but the parses available through the system are much impoverished. For example, annotations for logical subjects are omitted. Running Minipar through GATE is also at least an order of magnitude slower than as a stand alone application.

[152] See footnote on page 70 for an explanation of precision and recall.

[153] Named entity information has been omitted from these examples for clarity. Each line of Minipar output corresponds to a node in the dependency tree of the sentence. The fields in a typical node are:
NodeID (Token Lemma POS ParentNodeID DependencyRelation (gov ParentLemma) (Antecedent))

However subject and object control were not always correctly differentiated (presumably due to information lacking in the lexicon). For example the object control verb *expect* was correctly analysed, with *him* recognised as the canonical subject of the embedded verb *do* (3.136), and *promise* was analysed as a subject control verb (3.137) (*I* is the subject of *do*). However *asked* was analysed as a subject control verb, as in (3.136a), when it is in fact an object control verb and the correct parse of the dependencies of the verb *do* would be (3.136b).

(3.136) I expected **him**$_{\text{subj}}$ to **do** it

(3.137) **I**$_{\text{subj}}$ promised him to **do** it

(3.138) a. %**I**$_{\text{subj}}$ asked him to **do** it[154]

b. I asked **him**$_{\text{subj}}$ to **do** it

Another case where parsing failed, was if the direct object of a dative double construction was passivised, leaving a bare indirect object after the verb. This was incorrectly parsed, as in (3.139a), though this is not surprising, as the construction is very rare, and arguably only marginally well-formed.

(3.139) a. %It$_{\text{iobj}}$ was given him$_{\text{dobj}}$ by me$_{\text{subj}}$

b. It$_{\text{dobj}}$ was given him$_{\text{iobj}}$ by me$_{\text{subj}}$

Problems were also found with the correct attachement of prepositional phrases. Many NounPhrase–PrepositionalPhrase sequences that follow a verb are structurally and semantically ambiguous: the prepositional phrase may properly attach to the verb (as a verbal argument), or to an intervening noun phrase (as a noun phrase modifier). When a Preposition–NounPhrase sequence directly follows a lexical verb, the preposition may either be a phrasal verb particle, or be introducing a prepositional phrase.

For prepositional phrase attachment Minipar uses two principles (Lin Dekang, 1998). Minimal Attachment (see e.g. Hobbs and Bear, 1990) attempts to connect items in the least general (or alternatively most marked or collocated) way possible. For example in (3.140) the preposition *for* has a special syntagmatic relationship with verb *look*, meaning *search*. The sentence should be parsed with *look for* as a unit (3.140a). Though Minipar is heavily lexicalised (its word lists are based on WordNet), *look for* is not an entry, so this sentence is parsed as (3.140b).[155] This may yield an incorrect interpretation in which John is looking at or for something, on behalf of Mary.

(3.140) a. John [looked for]$_{\text{V}}$ [Mary]$_{\text{NP}}$

b. %John [looked]$_{\text{V}}$ [for Mary]$_{\text{PP}}$

The second principle of Right Association tries to attach constituents as far to the right (and so, in right branching languages like English, as low) as possible in the existing partial parse tree. According to Hobbs and Bear (1990, p.4) the Right Association principle should be subordinate to the Minimal Attachment principle. Hence, in a sentence like (3.141), where there is no privileged (marked) relationship between the preposition *in* and either the verb *phone* and or the noun *man*, the

[154] % is used here as a mark for an incorrect parse. See page 183 of the appendices for a list of notations used.

[155] The *look* phrasal units listed in the Minipar lexicon are: *look about, look across, look after, look around, look at, look away, look back, look backward, look like, look out on, look out over, look over, look through, look to, look towards*, and *look upon*.

in prepositional phrase attaches to the rightmost (and so lowest) point: the noun phrase headed by *man*. Similarly, where both verb and noun have a privileged relationship to the preposition (3.142), the rightmost attachment point is chosen again (*by* has the passive subject interpretation relative to *give* and the authorship interpretation relative to *book*). For such examples, Minipar behaves as expected.

(3.141) John phoned [a man [in Chicago]$_{PP}$]$_{NP}$

(3.142) John was given [a book [by a famous professor]$_{PP}$]$_{NP}$

However, in cases where the preposition is marked relative to the verb only (as *from* is to *buy* in 3.143a), and so should be attached at the verb phrase, Minipar does the reverse (3.143b). It appears that Minipar uses the principles roughly as they are described by Hobbs and Bear, but that the Right Association principle overrides the Minimal Attachment principle, and not vice versa.

(3.143) a. John bought [a book]$_{NP}$ [from Mary]$_{PP}$

b. % John bought [a book [from Mary]$_{PP}$]$_{NP}$

In cases where either reading seems appropriate (3.144), Minipar uniformly selects the right associated parse (3.144b).

(3.144) a. John made [coffee]$_{NP}$ [for Mary]$_{PP}$

b. John made [coffee [for Mary]$_{PP}$]$_{NP}$

Only in cases where the intervening noun is a pronoun, and so is less likely to be modified by a prepositional phrase, does Minipar choose the minimally attached parse (3.145a) over the right associated one (3.145b).

(3.145) a. She reported [him]$_{NP}$ [to the police]$_{PP}$

b. She reported [John [to the police]$_{PP}$]$_{NP}$

Since indirect objects are of particular interest to this study, the following heuristic as used to modify the Minipar output: prepositional phrases with *for* or *to* and that follow a bare noun phrase object were uniformly treated as indirect objects; other prepositional phrases in the same position were treated as obliques.

In the following analysis only arguments of full verbs were considered – the copula *be* and modal and auxiliary verbs were excluded (*will, have, might, could* etc.). In total 2.4 million argument phrases of 1.6 million verbs were extracted from approximately 15 million words of text.

3.3.3.2 Automatic Coding

Positive affect of whole arguments was calculated using the Dictionary of Affect in English (Whissell, 1989). The dictionary consists of more than 8000 word entries that have been rated by human participants for pleasantness, activation and imagery on a scale of 1 to 3.[156] When analysing the positive affect of an argument phrase, all its words were considered, and a mean affect value calculated. Words that did not appear in the dictionary were excluded from the analysis.

[156] On average each word received eight judgements, and the mean judgement across words in the dictionary is 1.84.

Arguments were considered to be definite if they were either a pronoun or proper noun (according to the parser's named entity recognition function), or if they were modified by the definite determiner (*the*), or were modified or consist of either demonstrative article or a possessive pronoun.[157]

For the ontological properties of arguments, a combination of lexical information, named entity recognition data from Minipar and information from the semantic network lexicon WordNet 2.1 (Miller et al., 1990) were used. The "humanness" of noun phrase referents was first checked by looking for pronouns – any argument whose head was a personal pronoun[158] was deemed to represent a human referent, and given a "humanness" value of 1. If the head noun was a proper name (as recognised by Minipar) it was deemed to be human if it fell under either of the Minipar entity categories of *person* (named people, such as *Tom*, or *Ms. Smith*) or *post* (names of professional roles, such as *the Prime Minister*), and did not fall into other non-personal entity categories, such as names of locations or organisations (consider *Paris* which is both a given name and a place name, or *Hoover* which is both a surname and a company name). For common nouns WordNet was consulted to determine the probability that the lemma in question represents a human concept. For example *collector* has senses that are descendants of the concept *person*, and so are animate (e.g. *rent collector*), and senses that are not (e.g. *rain collector*). Since WordNet has sense frequency data, it is possible to estimate the probability that an arbitrary instance of a word refers to a human or non-human concept. In the case of *collector*, uses of the person sense outnumber others, giving a "humanness" probability of 0.87. Similarly, a *carrier* may be human (e.g. of a disease) or an artefact (e.g. of aircraft). In this case inanimate uses predominated in the WordNet sense corpus, and the score given was 0.32.[159] Common nouns that did not appear in WordNet were excluded from the analysis.

Whether arguments referred to social groupings and to concrete objects was determined in a similar fashion, by consulting named entity information first, and then by looking at WordNet.[160] A general measure of animacy was derived from the measures for humans, social grouping or animals (insects and higher animals were included), by combining their probabilities productively.[161]

3.3.4 Results

For each grammatical, canonical and serial position measures of association were calculated against each of the structural features of interest. Each clause position can be treated as a binary variable. For example, to calculate the association between definiteness and the direct object, all direct objects in the corpus can coded as '1' on an "position" axis, and all other grammatical arguments (subjects, indirect objects and obliques) coded as '0'; definiteness of each of these arguments can be coded '1' or '0' on a "feature" axis. By taking a linear correlation coefficient over these two axes, the

[157] Demonstrative articles: *this, that, those, these, both, neither, either* considered; possessive pronouns considered: *my, your, her, his, our, their, mine, yours, hers, ours, theirs*.

[158] The list used was: *he, her, him, she, herself, himself, I, me, myself, oneself, ourself, ourselves, thee, thou, thyself, thineself, us, we, who, whoever, whom, whomever, whomsoever, whosoever, you, yourself, yourselves*.

[159] These probabilities were precompiled from the WordNet dictionary files. To allow for unattested word senses, a smoothing factor of 0.5 was added to the frequency count for each sense. As a result, for unattested words, the probability was determined by the number of human and non-human senses. For example *hammerhead* has a human, an animal and an artefact sense, none of which are attested with corpus examples in WordNet. It was given a probability of 0.33 for humanness. Such an approach was judged better to automatic word-sense disambiguation methods, which remain computationally intensive and error-prone. For example, a state of the art automatic sense annotator such as SenseLearner (Mihalcea and Csomai, 2005) achieves accuracy levels just under 70% on WordNet senses. However a naive baseline, in which the most common WordNet sense is chosen for a token without reference to its context achieves 60-65% accuracy (*ibid* p.4).

[160] It would be interesting to investigate the association between concreteness of the verb, and active/passive choice, but WordNet does not distinguish between concrete and abstract verbs.

[161] $p_{anim} = 1 - \Pi(1-p)$. For example the word *chick* has a 0.8 probability of being an animal, and a 0.2 probability of being a person, and zero probability of being a social grouping, so the probability of it representing an animate entity is 0.84.

Table 3.27: Correlations between selected features and argument positions in the British National Corpus

	Grammatical Function			Canonical Function			Serial Position		
	Subj.	**Dir. Obj.**	**Ind. Obj.**	**Subj.**	**Dir. Obj.**	**Ind. Obj.**	**First**	**Second**	**Third**
n	972,250	885,973	39,035	1,247,870	1,002,290	162,007	1,536,199	829,830	52,326
Animacy	**0.42**	**-0.36**	0.01	0.41	-0.36	-0.10	0.26	-0.24	-0.08
Concrete	0.07	-0.06	-0.02	0.08	-0.07	-0.03	0.06	-0.05	-0.02
Definite	**0.29**	**-0.17**	0.02	0.23	-0.15	-0.15	0.18	-0.15	-0.10
Affect	**0.20**	**-0.18**	0.00	0.18	-0.16	-0.03	0.08	-0.08	-0.01
Person	**-0.30**	**0.25**	0.00	-0.26	0.23	0.06	-0.15	0.14	0.04
Length	**-0.32**	**0.31**	0.00	-0.20	0.20	-0.01	-0.10	0.11	0.00

strength of association is quantified. The other features of interest are numerical: probabilities (animacy, concrete), real numbers (affect) or natural numbers (person, length); and they can be similarly compared to a binary-coded clause position. In fact, the calculation was made simpler by using the point-biserial correlation, which is equivalent to taking a linear correlation with one binary axis, and is based on means for each binary class, and their relative proportions.[162]

The results are shown in table 3.27. The number of arguments sampled for each position are indicated in the row headed *n*. The figures are correlation coefficients *r*.

The strongest positive correlation of 0.42 found was between animacy and grammatical subjecthood as opposed to other grammatical functions,[163] and squaring this value yields the correlation of determination of 0.18. This means that just under 20% of the mutual patterning between animacy and subjecthood can be explained. The correlation of -0.36 with the grammatical direct object indicates that, all other things being equal, a direct object is more likely to be inanimate, or conversely, an animate argument is more likely to appear in a position other than the direct object.

Animacy patterns strongly with both grammatical and canonical subjects (as predicted by the identification of subject with participant roles such as Agent and Experiencer), and with the first serial position (as predicted by anthropocentric saliency). While direct objects are associated with inanimacy, the grammatical indirect object has virtually no association with this semantic feature, and the canonical indirect object has a negative association.

Concreteness showed very light correlations, and no effect of saliency on serial position can be seen.[164] The relative ubiquity of concrete arguments may account for the small correlations. That is, since abstract arguments (as indicated by WordNet) are rather rare any difference in their distribution may be drowned out by concrete arguments.

The structural features that were expected to pattern predominantly with serial position, had in fact stronger associations with grammatical or canonical positions. Person correlates considerably with serial position, but has a stronger relationship with grammatical and canonical function. Affect was most strongly associated with grammatical subjecthood. Similarly, definite arguments tend to

[162] The point-biserial coefficient measures the association between numerical data and a dichotomous binary variable. It is equivalent to the Pearson coefficient, in which one of the variables has only two values. It is calculated on the basis of the mean of the numerical variable for each of the binary categories, and the relative proportion of cases that belong to each of these categories:

$r_{pb} = \frac{(\bar{x}_0 - \bar{x}_1)\sqrt{pq}}{s_x}$, where p and q are proportion of 0 and 1 values (so $p = 1 - q$). pq can also be expressed as $\sqrt{\frac{n_0 n_1}{n(n-1)}}$, $n = n_0 + n_1$

[163] The human, social grouping and animal features all had similar weaker correlations than this aggregate feature, so they are omitted from the table. For example the correlations with grammatical subject and object respectively were: human, 0.40 and -0.36; social 0.06 and -0.05; animal 0.009 and -0.002.

[164] Because of the very large sample sizes, such marginal correlations are nevertheless significant at conventional levels.

Table 3.28: Correlations between selected features of common nouns and argument positions in the British National Corpus

	Grammatical Function			Canonical Function			Serial Position		
	Subj.	**Dir. Obj.**	**Ind. Obj.**	**Subj.**	**Dir. Obj.**	**Ind. Obj.**	**First**	**Second**	**Third**
n	412,393	691,127	19,632	571,488	773,566	124,049	818,174	608,747	44,974
Animacy	0.18	-0.16	0.01	**0.24**	**-0.20**	-0.06	0.16	-0.14	-0.05
Concrete	0.07	-0.06	0.01	0.09	-0.07	-0.02	0.06	-0.05	-0.03
Definite	0.10	0.03	0.00	0.03	0.05	-0.15	0.07	-0.04	-0.08
Affect	0.02	-0.02	0.01	0.02	-0.02	0.01	0.00	0.00	0.01
Length	-0.24	0.24	0.02	-0.07	0.10	-0.06	-0.02	0.03	-0.04

precede indefinite ones, as is predicted by the identification of definiteness with given information, and indefiniteness with new information, but the correlations are strongest for grammatical positions.

In constituent length there is a strong asymmetry between the first and second argument positions, with shorter elements appearing first, and no association with the third position (one would expect this to have a stronger correlation than the second position). Again there is a peculiar but definite association between length and canonical position.

Associations were also calculated for common nouns only, as a large proportion of arguments seen in the corpus (ca. 40%) were proper nouns and pronouns, and these might bias results. These figures (table 3.28) do explain some of the puzzling patterns seen above. The patterning of positive affect seems to be completely determined by pronouns and proper nouns, which in the main describe people – as might be expected, subjects in affect experiments are more likely to be positively disposed to words that refer to humans. When this effect is removed, the correlation with canonical function disappears.

The association of definiteness and canonical position also disappears when only common nouns are considered, and a slight correlation remains with serial position.

The lack of a relationship between constituent length and serial position is hard to explain. As with the lack of effect for concreteness, it may be that an overwhelming majority of short argument phrases are drowning out the effect of longer ones.

Turning to verb features, there was a considerable negative correlation found between the passive voice and the progressive (i.e. gerundive or continuous aspect). Assuming that the passive expresses telic aspect and resulting states, as was suggested by the literature (see section 3.2.1.2), this is to be expected, as the progressive describes ongoing processes. Finally, a marginal association of 0.09 was found between negative affect and the passive form, indicating that there is a very slight adversative tendency in the English passive. However, these figures related to the passive should be treated a little cautiously, since an automatic parser like Minipar cannot distinguish between the dynamic and stative passives, which in English share the same structural form, unlike in German. Hence some of the sentences treated as passive constructions in this analysis may be copulative constructions (see page 74 for a description of dynamic and static passives).

3.3.5 Discussion

The patterning of animacy broadly followed expectations. Canonical subjects, which are associated with semantic roles that are typically animate (such as Agent and Experiencer), were more likely to be animate; canonical direct objects (associated with Patient, Percept, Theme) were less likely

to be animate. The first sentence position was more likely to be animate and second position less likely, as is predicted by notions of anthropocentric saliency. The fact that the effect was stronger for canonical positions than for grammatical positions (which confound canonical and serial positions) suggest that these two effects are at least somewhat independent.

The indirect object is associated with animate roles such as Experiencer, Recipient and Beneficiary, so it was a surprise that a strong positive association was not found with this position. Both the first noun phrase in a double object construction, and the object of *for* and *to* prepositional phrases following a direct object were included in this category, so a possible explanation is that sentences such as (3.146) were misparsed, treating *dinner* as an indirect object.

(3.146) %Terry came home$_{\text{dobj}}$ for dinner$_{\text{iobj}}$

Another possible reason is that the first noun phrase of DirectObject–Predicate sentences such as (3.147) are annotated by Minipar in the same way as indirect objects. And manual examination of the parser output also confirmed some cases where a noun-noun compound was incorrectly parsed as a double object construction (3.148).[165]

(3.147) The structure of agriculture **makes it**$_{\text{dobj}}$ **difficult**$_{\text{pred}}$ to treat the peasant drive for noble land as the harbinger of rural capitalism.[166]

(3.148) %Ignoring attempts on his life, and constant threats, Wallenberg used every means possible to outwit the executioners, ... **issuing Swedish**$_{\text{iobj}}$ **passports**$_{\text{dobj}}$ which, surprisingly, offered some protection in a period hardly renowned for respect for the law.[167]

The patterning for concreteness was not inconsistent with the idea that more tangible entities prefer to appear early in the sentence, but the effects were so slight that no conclusions can be drawn, particularly on the relative associations with serial, grammatical or canonical positions.

Finally, one would not expect that participants that are described in the 1st or 2nd person are more likely to be agents, as is suggested by the association figures, since person is a viewpoint-dependent feature, and agentivity is a feature of the nature of involvement in the situation described.

However, this pattern may be saliency driven choices when alternative verb lexicalisations are available. For example, in a symmetrical situation of similarity of appearance, in which the listener is a participant, the speaker is more likely to place the 2nd person argument first (3.149a) than second (3.149b).

(3.149) a. You look like Nikita Khrushchev

b. Nikita Khrushchev looks like you

Similarly in a transfer situation, a Recipient speaker may be more likely to choose the verb *got* or *received* (that allows them to put themselves in preverb position) than *gave*.

In summary, the corpus validation exercises provided qualified support for only some of the conclusions of the judgement experiments. Animacy is indeed associated particularly with subjects (which is consistent with both agency, and speaker saliency).

[165]The Minipar codes for IndirectObject–DirectObject sequences of *obj1–obj2*, is also used for DirectObject–PredicativeComplement sequences.

[166]BNC text FB1: *Rethinking the Russian revolution*. Edward Acton, 1992.

[167]BNC text A4X: *Independent Newspaper*, 10th of December 1989.

3.4 Expert Annotation of Roles

3.4.1 Background

Before Dowty (1991), semantic role accounts usually assumed a finite set of atomic (and often primitive) categories. For such a set of roles to be practically useful, it must be possible for different judges at different times to reliably duplicate each other's annotations to a large degree.[168] This section describes an annotation exercise undertaken with participants who are linguists. It asks judges to categorise each argument of a set of arbitrary English sentences as falling into one of several conventional semantic role classes.[169]

The degree of agreement seen between annotators can be compared to the amount of agreement individual judges showed when they repeated the exercise. Particular items can be compared by the level of consensus they elicited among judges, and each role category is compared to see the extent to what extent it overlaps with other categories. This provides suggestions on what sets of generalised roles (e.g. Dowty, 1991; van Valin and LaPolla, 1997; Primus, 1999) might be more successful, and extrapolations can be made from the data to estimate what levels of agreement would be found if such reduced inventories were used in such a task.

3.4.2 Materials

An arbitrary set of sentences for role annotation was collected. The materials came from two sources. Nine verbs were randomly selected from a list of verbs included in the British National Corpus; three were of high frequency; three of medium and three of low frequency (the verb list was partitioned into three part of equal length). The verbs selected were (in decreasing order of frequency): *say, know, take, play, let, decide, review, monitor, terminate*.

For each of these verbs, sets of random matching web search results (gathered from Google, using a random offset from the first result) were examined by hand until three appropriate examples of each could be found. Appropriate examples were those finite verbs with two or more arguments (including oblique phrases such as adverbials). For ease of interpretation for a general audience, examples with technical or ambiguous language were excluded (e.g. 3.150).

(3.150) **You can let the server guess**, select to force the use of ch.pgp.net, or choose from: at, ca, ch, dk, de, no, uk ...

Clauses with idioms and phrasal verbs were also excluded as their verb-arguments semantics are not generally compositional (e.g. 3.151).

(3.151) The media have a **crucial role to play** in keeping the tsunami in the spotlight

[168] To my knowledge the only published work that discusses the reliability of semantic role annotations of text is Blatt et al. (1984, pp.6-13,24-32). The authors describe a a planned semantic role annotation test among several judges, but difficulties in arriving at a consensus on an annotation standard resulted in the exercise being abandoned.

[169] A follow-on experiment was designed to try to elicit similar judgements indirectly from non-linguists. Test item arguments were compared to five reference arguments that represented different semantic roles. Participants were asked to choose which of these reference arguments was most like the test item argument. Participants had been instructed to attend to verb-argument meaning relationships when making these similarity judgements, by means of an calibration exercise that highlighted such relationships by illustration only – no linguistic concepts were introduced. However the results of a pilot with 9 non-linguist respondents were disappointing. Only 45% of test items were correctly identified as similar to the test items with which they shared a semantic role category. Since chance agreement is 20%, and expected agreement if subjects used grammatical role to identify similarity is 27%, the experiment was judged to be a failure.

So that participants would only have to attend to one verb-argument relationship, sentences in which a single constituent was governed by more than one verb were also avoided (e.g. arguments of raising and control verbs).

In addition to these 27 sentences, another 54 were taken from the English diathesis acceptability experiment (described in sections 3.2.2 and 3.2.3). These were the test items from that experiment, consisting of 30 randomly selected monotransitive active sentences, and 24 randomly selected sentences of the form Verb–NounPhrase–PrepositionalPhrase with *to* or *for*. All materials were gathered with two sentences of surrounding context, ideally one sentence before and after the test item. Each of the 81 test items was then segmented by hand into individual arguments, yielding a total of 219 constituents to be annotated. As discussed in the introduction (section 1.1) as few assumptions as possible were made about syntactic representations. All dependent clauses were treated as arguments, including both nominal and clausal adverbials.

3.4.3 Procedure

Practitioners and students of linguistics were recruited via a posting to the LinguistList mailing list.[170] Candidates were invited to take part in a web survey in which they would have to annotate text for semantic roles. The test items were presented on a single web page, in a single randomised order that was fixed across participants. The role labels are given in (3.152).

(3.152) Agent (or Actor), Patient, Experiencer, Percept (or Stimulus), Instrument, Recipient, Beneficiary, Theme, Manner, Time, Place, Reason

In the instructions, no explanation or description of the labels was supplied, as the objective was to study their existing interpretation among the linguistics community. The instructions also stated that the finer distinctions that can be made within roles were not of interest in the experiment – e.g. the Agent/Actor distinction; different types of locational and temporal roles; or various kinds of contingencies such as Condition, Reason, Purpose, Motivation. A practice item was supplied as part of the instructions. The instructions are reproduced on page 201 of the appendices.

For each test argument, the relevant clause was highlighted in its context (by underlining), and presented with the title of the document or film that it came from (film and scene descriptions were not given). A drop-down menu allowed the respondent to choose one of the role labels supplied, or to specify "other/unsure" for each argument. A text field was also given so that respondents could make comments for each if needed. Respondents choosing "other/unsure" were asked to use the comment field to suggest an alternative label, or otherwise to make a note of their difficulty with that item.

Respondents were also asked to supply their name and contact email, and to specify their dialect, and whether they are native speakers of English. Once all annotations were made, participants had to press a button at the end of the page to submit their annotation. Participants could end the experiment at any time by closing their browser, or submit partial annotations by jumping ahead to the submit button.

Several weeks after the experiment was conducted, respondents who supplied an email were again canvassed, and asked to repeat the experiment in whole or in part to test intra-subject agreement.

[170] A community website for Linguists [http://linguist.emich.edu/], viewed August 2006.

Table 3.29: Overview of responses for annotation exercise

Role	Votes	Subj.	Comp.	Items	Plur.	Pair.
Agent	784	778	6	56	94%	90%
Beneficiary	151	2	149	10	84%	74%
Experiencer	160	136	24	13	69%	55%
Instrument	65	27	38	4	48%	28%
Manner	52	-	52	4	67%	50%
Patient	445	20	425	25	62%	46%
Percept	90	14	76	3	49%	33%
Place	203	1	202	16	86%	77%
Reason	167	3	164	14	80%	70%
Recipient	86	8	78	6	59%	41%
Theme	423	45	378	32	54%	37%
Time	248	4	244	18	98%	91%
other	128	14	114	4	46%	32%
tied				14	38%	28%

Total votes for each role category are listed, and split by grammatical role (subject or complement)
The number of arguments to receive a plurality vote of each label are listed in the "Items" column.
The plurality and pairwise agreement appear in the last two columns.

3.4.4 Results

Overall, 21 people took part, 4 of those anonymously. Of those who specified, 8 were native English speakers and 10 were non-natives. Their dialects were: US (9), English (2), Irish (2), Australian (2) and "other" (2) varieties of English. On average each respondent made 142 argument annotations, out of a total of 219. Eleven of the respondents completed the survey in full. After being canvassed by email some weeks later six respondents made repeat submissions.

For each experimental item, a consensus annotation was determined – that is the mode or plurality category, that gained more responses than any other across participants (in most cases this was a majority judgement). Of the 219 arguments, four were judged "other/unsure" by a plurality of judges, and 14 resulted in ties (nine or which were Patient/Theme ties). The degree of agreement on a single role ranged from 100% to 22%. The worst performer (i.e. that on which there was least consensus) was the underlined subject of the sentence below.

(3.153) "<u>your years with us here at Rydell</u> have prepared you for the challenges you face"[171]

In the first three data columns of table 3.29, the total number of annotations is given for each role label, and this figure is split by annotations for subject arguments, and complement (i.e. non-subject) arguments. For some role labels, there were clear preferences for subjects or complements: for example Agent was ascribed almost exclusively to subjects, and Place almost exclusively to complements.

The second half of table 3.29 is based on consensus annotations: the number of items (among 219) to receive that role label from a plurality of judges; the average plurality figure over those items; and the average proportion of pairwise agreement among judges (i.e. for a single item, the number of pairwise agreements among judges, over the number of pairs of judges – see section 3.1.2.4).[172] The average consensus agreement was 74% – that is, for an average argument, three-quarters of

[171] *Grease*, Randal Kleiser, 1978.

[172] While the pairwise agreement figures are uniformly lower, they correlate very highly with the plurality figures; $r_p = 0.982,\ p < 0.001$.

Table 3.30: Confusion figures for annotation exercise

	n	Ag	Bnf	Exp	Instr	Man	Pat	Prcpt	Place	Reas	Rcpt	Thm	Tm	other
Agent	56	*94%*	<1%	3%	<1%	-	<1%	<1%	<1%	-	1%	<1%	-	1%
Beneficiary	10	-	*82%*	1%	-	1%	3%	-	-	3%	8%	1%	-	2%
Experiencer	13	**17%**	1%	*71%*	2%	-	3%	1%	-	-	1%	4%	-	1%
Instrument	4	**15%**	-	-	*46%*	8%	2%	6%	-	4%	-	8%	-	12%
Manner	4	4%	2%	-	**17%**	*67%*	-	-	-	-	-	-	-	10%
Patient	25	<1%	1%	3%	1%	-	*61%*	1%	-	-	1%	**30%**	-	1%
Percept	3	-	-	-	-	-	**32%**	*49%*	-	-	2%	**17%**	-	-
Place	16	-	1%	-	-	1%	1%	-	*87%*	-	3%	-	-	7%
Reason	14	1%	1%	-	1%	5%	-	1%	-	*79%*	1%	1%	-	11%
Recipient	6	-	9%	2%	-	-	5%	-	**16%**	-	*59%*	4%	-	6%
Theme	32	-	-	<1%	3%	<1%	**29%**	7%	<1%	1%	<1%	*54%*	-	5%
Time	18	-	-	-	-	1%	1%	-	-	<1%	-	1%	*95%*	2%
other	4	2%	8%	7%	-	-	8%	-	-	**21%**	2%	7%	-	*46%*
tied	14	2%	-	1%	4%	-	**34%**	**15%**	4%	1%	1%	**33%**	2%	5%

Rows group votes by the consensus category given to each item

Columns describe proportion of votes given to those items (column labels are abbreviations of row labels)

linguist respondents could agree on a single role as the correct annotation. Subject arguments were on average more reliable (average 85% agreement) than other arguments (mean 68%).

The plurality and pairwise measures can also be adjusted for what degree of agreement might be expected by chance. Taking the distribution of roles found in all responses to estimate their distribution in the language at large (which is a conservative assumption, relative to the assumption that all roles are equally probable), agreement measures can be normalised to a scale where 1 signifies full agreement, and 0 signifies only the degree of agreement predicted by chance (see formulae 3.3 and 3.7, page 69). By this normalisation the average plurality agreement is 0.65. The same measure calculated separately by grammatical function was 0.44 for subjects, and 0.59 for other types of arguments. Using the pairwise measure, the normalised overall agreement was 0.57 (the Kappa statistic, $p < 10^{-12}$; see formulae 3.6 and 3.5). So, by various assumptions and measures, the level of agreement is somewhat more than halfway between what one would expect purely by chance, and full agreement.

To evaluate to what extent the role categories are confused with each other by judges, a confusion matrix was constructed (table 3.30). Each line of the table describes a single role, and the argument items that received it as consensus judgement. The columns then show what votes that set of items received. For example, among the votes given to the three items that a plurality of judges annotated as a Percept, just under the majority were for that category, but Theme and Patient also received considerable proportions.

The responses of individual participants can also be compared with the overall consensus. The extent to which individual respondents agreed with the consensus view varied, from 54% to 80%. On average, natives agreed with the consensus view 73% of the time, slightly more than non-natives (68%).[173]

In the follow-on intra-subject experiment, the six repeat-respondents agreed with their original annotations an average 79% of the time, ranging from 63% to 99%.

Despite the length of the experiment (which many of the respondents commented negatively

[173] Not significant: $p = 0.16$, 2-tailed independent samples Welch t-test over respondent sessions.

on), there was a minimal order effect. One might expect respondents to become less careful as they proceed through the exercise, and agreement to decrease. However, only a very slight correlation was found between order and the plurality agreement measure ($r = -0.08,\ p = 0.24$).

3.4.5 Discussion

The results can be described in qualitative terms with reference to the consensus view, and the degree of confusion with other categories. Argument items that were judged to represent an Agent by the plurality of respondents were very stable, with a mean plurality agreement of 94% (i.e. for argument items that a majority of participants thought were Agents, on average only about one in twenty respondents thought the item represented a different role). Agents of cognitive actions with little or no effect on objects were sometimes annotated as Experiencer (e.g. with the verbs *play*, *find*, and *seek*), and Agents whose actions entailed possession got some votes of Recipient (*get, take*). Similarly items that were judged to represent Experiencers, and might involve active participation or volition gathered quite a few votes of Agent (e.g. the verbs *know, hate, want, see*). Annotators were reluctant to give inanimate referents (particularly clausal constituents) the role of Agent, despite the fact that the role was explicitly conflated with the Actor role in the instructions, often opting instead for Instrument, or suggesting an additional category of Cause. Some annotators commented that it would be desirable to distinguish between active and passive perception (i.e. *listen* versus *hear,* and *look* versus *see*). Recipients were often Beneficiaries also, and so there was some disagreement between these two roles (e.g. *you* in each of these three cases is a beneficial recipient: *compensate you*; *bring your son back to you*; *fetch slippers for you*).

The Patient/Theme distinction seemed to be the largest source of disagreement and confusion among judges, and accounted for a majority of the 14 items in which there was a tie among annotators. Prototypically affected participants tended to be labelled Patient (e.g. *man* in *Dinosaurs eat man*), with a cline of decreasing affectedness towards prototypical Themes (e.g. *take orders*). In addition, judges preferred concretely affected and sentient Patients (i.e. Patients who are also Experiencers). Inanimate, abstract or event participants were more often annotated as Theme. Several judges also suggested an additional Product role, for participants that are created during an event. Percept was used very little, although there were plenty of candidates to which this role might apply. Judges preferred Theme in these cases (e.g. *analyze their attack, find something*).

Despite the note in the instructions saying that different subtypes of contingencies were not of interest, many variations on Reason were suggested, including Cause (a directly precipitating event), Condition (a potentially limiting event), Purpose (motivation for an Agent) and Result (a precipitated event).

The Place and Time categories were highly reliable. However many judges noted that they would prefer to have more precise Source, Path and Goal labels for the spatial and temporal domains. Reflecting localist approaches (see section 2.4), these two categories were also occasionally suggested for non-spatial and non-temporal cases such as communication (e.g. Source for the subject of *say*). Range and Measure were also suggested as categories for amounts of time and money.

Some respondents also questioned what level of semantic analysis was appropriate (cf. section 1.1) – that is to what extent annotations should be made on the basis of literal, derived or pragmatically presumed meaning. For example in the sentence (3.154a), *what* can be considered a Cause on a literal reading, but is rather a Reason on the interpretation represented by (3.154b). Similarly *him* might be a Recipient on a literal reading of *they chose a new form for him*, but in its context

Table 3.31: Agreement measures over different role sets

	Items	Votes	Roles	Kappa	Plurality
All	219	3002	12	0.57	0.64
Nuclear Roles	150	1976	8	0.55	0.56
Dowty Proto-Roles	150	1976	2	0.86	0.92
Dowty *plus* Proto-Recipient	150	1976	3	0.85	0.90
Dowty *plus* ProtoExperiencer	150	1976	3	0.81	0.87

"kappa" is a global measure of pairwise agreement, normalised using the distribution of categories seen in the data.
"plurality" is a similarly normalised, but based on the proportion of agreement on the mode (i.e. most common) category.

the correct interpretation seemed to be *they changed him into a new life-form*, in which case Patient would be more appropriate.

(3.154) a. What brought you to Casablanca?[174]

b. Why did you come to Casablanca?

It is interesting to note that intra-subject agreement is only marginally higher than inter-subject agreement. This suggests that a considerable proportion the variation seen is due to participants not being sure or reliable in their responses, rather than being primarily due to differences between participants (which might either stem from participants having different understandings of the test items, or having differing interpretations of the linguistic role categories).

Given that judges were not in full agreement, one can also speculate on what might be a better set of roles to elicit more consensus. Two proto-roles of Agent and Patient have been suggested by Dowty (1991) (van Valin and LaPolla, 1997, propose the broadly identical "macroroles" of Actor and Undergoer). Primus (2002) adopts much of Dowty's approach, but also suggests a Proto-Recipient role, to deal with the linking of the dative case in German.[175] One can try to extrapolate from the data returned in this exercise to estimate what degree of consensus would have been found if the same participants had been presented with such a reduced set of roles.

Since the proto-role sets are intended to explain linking patterns for nuclear arguments only, more circumstantial roles were removed from this analysis. Annotations of the roles of Manner, Place, Time, Reason and Other were discarded, as were argument items that received a plurality vote of one of these categories, or were tied. This reduced set of data includes 149 argument items (out of an original 219), and 1974 responses using one of eight role categories (Agent, Beneficiary, Experiencer, Instrument, Patient, Percept, Recipient, Theme). The agreement measures were recalculated for this nuclear argument set, to provide a comparable baseline agreement figure (see the second line of table 3.31). As can be seen, both Kappa (pairwise) and plurality measures of global agreement fall marginally from the original set to the reduced set.

The first reduced set of roles to be tested was Dowty's, using the following reduction (the proportion of test items included in each classification is given in brackets):

(3.155)
- ProtoAgent: Agent, Experiencer and Instrument (49%)
- ProtoPatient: Patient, Percept, Theme, Recipient and Beneficiary (51%)

[174] *Casablanca,* Michael Curtiz, 1942.

[175] As discussed in section 2.6.1 (page 56), for Primus the Proto-Recipient role does not have an independent characterisation, rather being a secondary concept defined in relation to the ProtoAgent and ProtoPatient.

With this role set, the reduced annotation data yields much higher agreement measures – and these measures are normalised for expected agreement, so they they have offset the fact that agreement is easier to reach with a smaller set of categories. But as discussed previously (section 2.6), this reduced inventory has trouble describing many common argument structures, particularly those that involve indirect objects.

Primus (2002) addressed this shortcoming of Dowty's by adding a Proto-Recipient to the scheme. The data from this exercise was collapsed into three proto-roles as follows:

(3.156)
- ProtoAgent: Agent, Experiencer and Instrument (49%)
- ProtoPatient: Patient, Percept and Theme (39%)
- Proto-Recipient: Recipient and Beneficiary (11%)

By this partition of role categories, both pairwise and plurality-based measures are very close to those of Dowty's two proto-roles. This is in part due to the fact that there are very few Recipient or Beneficiary items among the materials (16 out of 149). And these two roles may be overrepresented in the data set (relative to the language at large), since about a quarter of the sentences had a dative or benefactive clausal structure.

An alternative partition that may be more balanced relative to the incidence of each proto-role in the language groups Recipient and Beneficiary together as follows:

(3.157)
- ProtoAgent: Agent and Instrument (42%)
- ProtoPatient: Patient, Percept and Theme (39%)
- ProtoExperiencer: Experiencer, Recipient and Beneficiary (19%)

This performs marginally worse than the previous two partitions, but is still much superior to that of the original traditional partition of roles.

Overall, the exercise described in this section demonstrated considerable agreement on a conventional inventory of semantic role labels over an arbitrary set of English sentences. Most disagreement centred on the distinctions between Theme-like (Theme, Patient and Percept) and Actor-like roles (Agent, Instrument and Experiencer). Extrapolation from the data suggests that collapsing conventional roles into macroroles or proto-roles would yield higher agreement between judges, not only in nominal terms, but relative to the amount of agreement expected for smaller numbers of categories, and so implies that this level of granularity better reflects the internal cognitive categories used by speakers.

3.5 Conclusion

The diathesis acceptability experiments in sections 3.2.2 to 3.2.5 demonstrated that such structures are not exclusively licensed at the lexical or structural level, as some theories would suggest. Features of semantics and discourse, both inherent to truth-conditional meaning and governed by choices of speaker perspective, can have consequences for the productivity of these constructions, and violations to these features can have similar consequences for acceptability to those seen for structural violations. The set of features seen to operate are consistent with the idea that elements of involvement, speaker saliency and information structure contribute in parallel to wellformedness.

These conclusions received qualified support from the corpus study 3.3, which found a weaker pattering of animacy with grammatical function, than with either serial position (consistent with saliency) or canonical function (consistent with participation based theories).

The role annotation study found considerable agreement among judges, but there was a lack of consensus on the exact boundaries among ProtoPatient like roles, and among ProtoAgent like roles. Several ways of collapsing conventional roles to generalised roles increased estimated agreement relative to chance expectation.

Chapter 4

Conclusion

The literature review in chapter 2 demonstrated that there remains a general lack of consensus on the fundamental sources of linking patterns. Some authors investigate the lexical semantics of individual verbs for explanations (e.g. Pinker, 1989; Rappaport and Levin, 1988), while others look to general principles in the grammar (e.g. Dowty, 1991; Goldberg, 1995). Some think that notions of pragmatics should be excluded from such accounts (Dowty, 1991), while for others they are crucial (Croft, 1991).

The empirical studies presented in chapter 3 indicated that properties of the semantics and pragmatics of situations can have an effect on the acceptability and productivity of particular argument structures. Furthermore, the effect of violations of such constraints were of a similar order of magnitude to those seen due to clear syntactic violations. And crucially these determining factors may be properties of the particular situation being described, rather than typical event properties, or entailed properties of verbs used to describe such an event.

This suggests a model of linking that does not assume primacy of one sort of constraint over another (say in the choice of the verb *buy* or *sell*, or of the active or passive voice), and so one that does not assume strict modularity of linguistic processing. Further, it suggests that all types of constraint be allowed to compete equally to select an optimal clausal expression.

In this chapter, the first section (4.1) reviews experimental evidence and theoretical accounts that argue against strictly modular accounts of language production and comprehension, and speculates on the various forms a non-modular process could take. Section 4.2 argues that competition between argument structures does appear to operate at the clausal level (rather than just at the level of the argument) in several languages; that these languages show effects of both semantic and pragmatic properties on linking; and that they broadly agree on the surface structural representation of typical transitive events. It is in the realisation of more balanced situations that subtle effects of participation and perspective emerge, and that differing syntactic requirements of each language result in varying argument structures.

Section 4.3 presents a possible reworking of Dowty's model of English argument structure realisation that is compatible with these findings, and serves as a proof of concept for such a parameterised non-modular competition-driven model. A selection of both semantic and discourse features are used. Their effects are combined using simple linear equations, based on relative weightings derived from the annotated materials returned in the exercise described in section 3.4. Candidate clause realisations are compared on the basis of their suitability to describe the situation in question, in terms of their similarity to a prototype transitive clause. The model is evaluated by examin-

ing its predictions on constructions that have received extensive attention in the linking literature (*like/please*, *buy/sell*, and the *load/spray* diathesis, among others), and the suitability measures it yields are correlated against the relative judged acceptability of dative/benefactive and passive sentences from the English language experiment described in sections 3.2.2 and 3.2.3. Despite its numerical simplicity, and the very small corpus of sentences used to derive its parameters, it provides correct predictions of the optimal clause realisation for the majority of phenomena examined, and correlates strongly with judgements of perceived relative acceptability.

Finally, in section 4.4, the implications of such a model are discussed, including speculation about the possibility of generalising it to other nominative and non-nominative languages.

4.1 Modularity and Language Processing

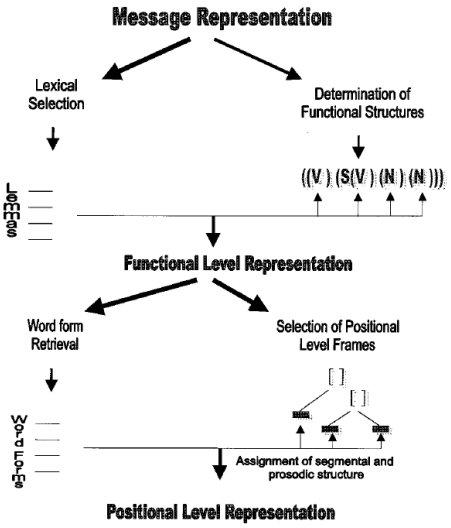
Various kinds of modularity (functional and/or informational encapsulation) have been posited for the human language function (Chomsky, 1986) or for the mind as a whole (Fodor, 1983). Detailed models of language processing have been constructed (e.g. Garrett, 1984; Bock and Levelt, 1994; Levelt, 2001) that assume distinct levels of linguistic function, whose interaction is limited, and which do not have direct access to each other's internal informational states. Linguistic theories such as LFG formalise such levels into independent levels of representation.

However, some experimental results have suggested that the language faculty is more intimately entwined with general cognition than such theories would presume. Not only have some semantic and syntactic violations of wellformedness been shown to cause similar effects in electrophysiological response (calling language-internal modularity into question – see page 11 in section 1.3), but many extra-linguistic effects have also appeared within similar time-frames (typically within the first half a second of comprehension). In both eye-tracking and EEG experiments, general background knowledge and knowledge of the current state of affairs in a discourse have been shown to effect online comprehension. Priming effects in picture-naming experiments, and studies of spontaneous and induced lexical errors, have both raised similar questions about language production, suggesting that syntactic, semantic and phonological processes are more tightly integrated than some believe.

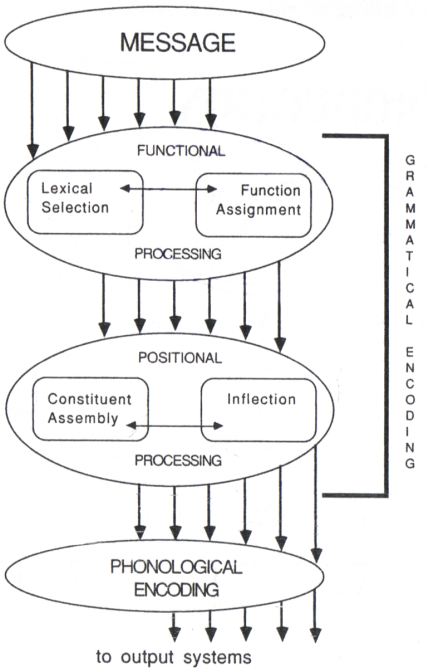
Tanenhaus et al. (1995) showed that the planning of eye-movements (saccades) can begin *before* the end of a word is reached during a listening task, and that precise timing is modulated by a visual scene presented. Participants were asked to follow simple instructions to manipulate objects in their visual field. For example, if both a candle and some candy were visible when a participant was given the instruction *pick up the candy*, eye-movements towards that object began 230 ms after the end of the word (saccadic planning takes about 200ms). However, if only one of those objects was present, the eye-movement started within 145 ms seconds of the end of the word, meaning that planning of the saccade began after only the first syllable /kæn/ had been processed. In a similar study, Karabanov et al. (2007) showed that no extra processing time was needed to resolve the reference of pronouns, relative to full noun phrases, suggesting that the (presumably extra-linguistic) operation of referent binding is well integrated into linguistic processing.

Hagoort et al. (2004) demonstrated that the N400 ERP component, usually associated with violations of language semantics, can also be elicited by general knowledge implausibilities. The violation in the utterance *The Dutch trains are sour* can be explained by a language-internal mismatch between the selectional restriction of the word *sour* for foodstuffs, and the semantic type of *train*. However, a similar effect was elicited for *The Dutch trains are white*, which must involve access to general knowledge (Dutch trains are all yellow). Özyürek et al. (2007) found that similar

Figure 4.1: Two modular views of sentence production



Garrett (1984)



Bock and Levelt (1994)

effects could be found when there was a semantic mismatch between a spoken utterance and an interspersed gesture, which involves integrating visual information (see Kita, 2000 and MacNeill, 2005 for corresponding theoretical claims on gesture).

Nieuwland and Berkum (2006) and Nieuwland et al. (2007) found that linguistic violation effects could be overridden in the presence of a supporting context. In the first study, sentences that usually elicited an N400 component, such as *The girl comforted the clock*, no longer did, given an appropriate context (e.g. a child's story in which the clock is a character). In the second, an ERP effect that is associated with referential binding ambiguity (e.g. referring to *the girl* when two girls have been mentioned in the discourse), did not appear if one of the potential referents had left the discourse (e.g. by dying). Together, the experiments just described suggest that both general semantic knowledge, and current episodic knowledge, are accessed online during language processing.

Vigliocco and Hartsuiker (2002) review both theoretical arguments and experimental evidence on the degree of modularity present in language production. Modular models often assume information encapsulation, unidirectionality and local interaction – that is, each module only receives the necessary information from modules with which it interacts (it does not have access to their internal states); information flows in one direction only (there is no capacity for feedback); and processing proceeds serially with interaction only taking place between neighbouring modules. Figure 4.1 illustrates two such models. As these show, such theories often assume both separation between lexical and phrasal operations; and between conceptual, grammatical, serial and phonological selection processes.

Early empirical support for modularity in lexical selection, between the retrieval of an appropriate lexeme (which is primarily meaning driven) and the selection of its form (which may involve

morphological processing) came from studies by Garrett (see e.g. 1984) into spontaneous lexical substitution errors. Such mistakes often involve some similarity between the intended word (the "target") and the uttered word (the "intruder"), and the similarities observed were either phonological (e.g. *shoulder* and *soldier*) or semantic (e.g. *shoulder* and *elbow*). Substitutions which had both meaning-based and form-based similarities were not found, suggesting that the corresponding processes are distinct (see Vigliocco and Hartsuiker, 2002, p.445).[1]

Evidence against such modularity can be seen in attested examples such as (4.1), where the intruder word *sleast* combines phonological elements of the semantically related *slightest* and *least*.[2]

(4.1) didn't bother me in the **sleast** ... slightest

A study of speech errors in German by Marx (1999) found that semantically related noun intruders shared grammatical gender with the correct target word more often than would be expected by chance (suggesting that the syntactic and semantic components interact), and that this effect could not be accounted for by correlations between gender and semantic similarity (Vigliocco and Hartsuiker, 2002, p.455).

Evidence from other experimental paradigms include Peterson and Savoy (1998). In priming experiments, they found that picture naming was speeded by preceding pictures which were phonologically or semantically related. Crucially, the phonology of unrealised lexical candidates had an effect. For example, naming *couch* speeded the naming of the subsequent stimulus *soda* despite the fact that the near synonym *sofa* had not been produced (strictly modular accounts would predict that only the selected lexeme – *couch* in this case – would have an effect on processing in other modules, such as the phonological one). Finally, there is extensive evidence of conceptual effects on agreement phenomena. Words that are grammatically singular but conceptually plural (such as the collective nouns discussed in relation to animacy in section 3.2.1.5) often take plural verb agreement in English (in British English this is regular; in American English such constructions are considered ungrammatical but are nevertheless more common than plural agreement with singular nouns that have conceptually singular referents). Error elicitation studies (see Vigliocco and Hartsuiker, 2002, p.455) in Italian have shown that nouns that have both grammatical and conceptual gender (such as *ragazza* 'girl') are given incorrect agreement with adjectives less often than words that only have grammatical gender (e.g. *panchina* 'bench' is also feminine).

The question of what information comes to bear in the generation of clause structure, and in what order remains controversial. Chomsky describes the process by which the generation of utterances are initiated ("the choice of action") as "largely mysterious", and he restricts himself to explaining how sentences come about once its constituent lexemes have been chosen (Chomsky, 1995, p.227). The problem is not inconsiderable, as the alternative expressions of a simple scene in (4.2) show (adapted from Fillmore, 1977b, p.103). There are many possible utterances to describe this event, even without considering the multiple lexicalisation possibilities of verb (e.g. *buy/purchase/acquire*; *sell/flog/peddle*; *pay/cough up/fork out*) and noun (e.g. *dollars/bucks/notes*; *puppy/little dog/pup/doggy/pooch*); optional reorderings of prepositional phrases; and less common constructions such as topicalisations and cleft sentences.

[1] However, since such substitution errors are very rare, the failure to find such errors may not be so surprising (according to Levelt, 2001, p.13464 such lexical selection errors occur no more than once or twice every 1000 words).

[2] Vigliocco and Hartsuiker (2002, p.452) credit this example to Boomer, D. S., & Laver, J. D. M. (1968). *Slips of the tongue*. British Journal of Disorders of Communication, 3, pp.2-12.

(4.2) a. Harry bought the puppy from Mr. Smith for $60

b. Harry bought the puppy from Mr. Smith with $60

c. The puppy was bought from Mr. Smith by Harry for $60

d. The puppy was bought from Mr. Smith by Harry with $60

e. Mr. Smith sold the puppy to Harry for $60

f. Mr. Smith sold Harry the puppy for $60

g. The puppy was sold to Harry by Mr. Smith for $60

h. Harry was sold the puppy by Mr. Smith for $60

i. Mr. Smith charged Harry $60 for the puppy

j. Harry was charged $60 for the puppy by Mr. Smith

k. $60 was charged to Harry for the puppy by Mr. Smith

l. Harry paid Mr. Smith $60 for the puppy

m. Mr. Smith was paid $60 for the puppy by Harry

n. $60 was paid to Mr. Smith for the puppy by Harry

There are indications that lexicalisation cannot strictly precede the consideration of final argument structure (as Chomsky suggests), since entity-specific factors, both ontological (e.g. animacy) and discourse related, (e.g. new/givenness) may interact with structural realisation possibilities to influence what part-of-speech a given element is realised as. For example, if two of the entities to appear in a given utterance are activities, either or both may become a verb or a noun, and the choice of one over the other is determined by (possibly conflicting) constraints on word order (e.g. for the purposes of text cohesion). In the example below, the choice of the election event or the opposition event as the verb appears to be driven by discourse considerations – that is which of *him* or *them* is given in the context, or carries rhetorical focus.

(4.3) a. He opposed their election to the council

b. They were elected to the council in the face of his opposition

If one does not assume that language production is modular, the space of possible processing models broadens greatly. While the evidence presented in sections 3.2.2-3.2.5 does not provide any insights into the sequential details of processing, it does suggest that many linguistic factors (which are traditionally ascribed to different modules) have similar effects on perceived acceptability, and that multiple candidate structures must be compared to arrive at an optimal utterance (as will be discussed further in the following section). It is nevertheless instructive to consider the broad space of possibilities.

A search model would be one that recursively investigates all possible sentence structures (e.g. passive/active, ditransitive and other diatheses) with each appropriate combination of candidate lexemes (e.g. {*opposed, election*}, {*fought, election*}, {*opposed, voting in*}, etc – only verb choice is mentioned here, but noun choice would multiply the possibilities).[3] The model would then have to rank all candidate realisations by some measure of acceptability, and choose the most suitable. Another model that would entertain all possible realisations is a race model. Several candidate "threads"

[3]Here the term "search" is used in the computational sense of a process (an algorithm) that iterates through the space of all possible solutions to a task, to find an optimal (or near optimal) answer.

(another computational analogy) would be initiated simultaneously, and the first to provide a solution that expresses all required elements, and passes some threshold of acceptability, would be chosen.

Such models can be considered uneconomical in terms of the cognitive load they require, since all solutions, no matter how unlikely, are attempted. Some have suggested that the initial choice in clause structuring may be made on the basis of discourse considerations. In Ferreira's (2000) subject-centric model, the topic of the sentence – a constituent that is given in the discourse, and so provides cohesion with what preceded it – is initially selected to be realised as the grammatical subject of the sentence. This then greatly reduces the range of possible realisations. For example in (4.3b) the choice of *they* as topic forces the clause describing the election event to precede the description of opposition to it. It also requires the passive voice, as there is no lexical verb in English with the meaning of "to be elected". In cases that the preferred subject is not compatible with an acceptable sentence that includes all the required elements, such a model might backtrack and select the next optimal subject around which to build the sentence. Anecdotal support for this model is provided by the fact that speakers sometimes repair half spoken sentences, to change their choice of subject. For example in (4.4) the speaker hesitates at the first *I mean*, chooses an expletive existential *there* as subject, and then repairs to a choice of *I* as subject, to admit the phrase *I should remind you* for rhetorical reasons.

(4.4) Do you, I wonder if you think your [*sic*] a good driver, **I mean there are, I mean I, I should remind you there are** five thousand people killed on Britain's roads every year, there are sixty three thousand people seriously injured, they reckon that car accidents cost something like five thousand million pounds in Britain every year.[4]

A verb-centric model is also possible – Jackendoff (1972, p.37) believes that "the verb of the sentence is what determines the relationship" to its arguments. In a similar manner to Ferreira's model, the verb can be chosen first, and then other elements slotted in around it if possible. If the best resulting candidate structure is suboptimal, a solution using an alternative verb can be attempted. For example, in (4.5) it appears that Sheila changes her mind about the choice of verb, causing a repair.

(4.5) **Sheila**: ... Mm ... well her was only thirty two Wendy, I don't know ... so she's only
Wendy: Well they just tell lies Sheila that's all
Sheila: she was on about it, I mean John hasn't got er, a bill and **Tony hasn't like, well he's not down**, but David's got a bill and he should be up court on Monday[5]

Vygotsky and others (see Jespersen, 1924, pp.147-150) have a more flexible view of initial lexical choice, in which the "psychological subject" (that is, the given element of the message) is "the first idea in [the speaker's] consciousness", while the "psychological predicate" is the new information to be added to the discourse and the "carrier of topical emphasis"[6] (Vygotsky, 1934, p.127). Vygotsky is speaking in the terms of the classical grammatical tradition, in which a sentence is composed of two parts: the subject, both a topical entity (i.e. the Theme) and grammatical unit; and the predicate, what is said about the subject (i.e. the Rheme) (Halliday, 1985, pp.30-33). In this view, the difference between the subject and verb-centric models would be that in the first the grammatical subject is the psychological subject, and in the second the verb is the psychological subject.

[4]BNC text FLA: Spoken discussion about transport among Scottish women.

[5]BNC text KR0: Recorded conversation between Sheila (52, North-East England) and Wendy (50, North-Central England). *To go down* means to be given a prison sentence.

[6]The term "topic" is used in many ways in the literature, confusingly referring to both Theme and Rheme (in Halliday's terminology).

Various views on the degree to which syntactic and semantic argument structures or frames are stored in the lexicon or generated on the fly can also be considered. A static model in which all frame information is encoded in the lexicon, specifying the semantic and syntactic correspondences for every argument, is not plausible in the light of novel uses (as was discussed in chapter 1). For example in a novel use of a noun as a verb (4.6, repeated from page 3), the speaker may not have prior knowledge of its argument structure. Presumably general principles are used to decide on an appropriate linking pattern for novel verbs, or perhaps the speaker makes an analogy with an existing verb that is semantically related, and adopts its structures (or calls on an inventory of standard constructions, such as those proposed by Fillmore and Kay, 1987; Goldberg, 1995).

(4.6) G.W. Dickinson, ex-superintendent of the Montana Union, has been gold-watched by the employees of the company.[7]

The contrary viewpoint, where no frame information is stored, and all argument structure correspondences are generated dynamically does not prove convincing. Despite the fact that semantic classes of verbs share much of their argument patterning, there are common exceptions – consider the verbs *break down/destroy/collapse* which were shown to have different syntactic frames in an identical context (see page 2). This model would also be very uneconomical in terms of processing load.

A model that occupies the middle ground might store exceptions only, and generate everything else on the fly. Perhaps only the common frames of common verbs are stored, for reasons of economy (striking a balance between the storage load of a large number of static frames; and the processing load of generating all frames dynamically). In any case the verbal lexicon does not seem to be an entirely static entity.

The evidence examined in this thesis cannot determine which, if any, of these models are the correct one. As a result no assumptions are made about modularity or processing sequence in clause generation. The model presented is "flat" in that none of the factors that help determine argument structure is given prior or privileged influence on the output. All possible realisations compete to find the most appropriate solution.

4.2 Competition in Clause Generation

As mentioned in the last section, the evidence from the diathesis experiments (3.2) suggested that many semantic and pragmatic factors co-determine perceived acceptability. If one further assumes that differences in acceptability are reflected in differences in productivity, this suggests that competition plays a role in the selection of appropriate argument structures.[8] Notions of competition are common in models of language processing, though not universal (see Mahon et al., 2007, for a recent dissenting view). La Heij (1988) proposed that lexical search is a competitive process, with activation levels driving selection (see also Levelt, 2001, p.13465, for a claim that lexemes are tailored to the listener's state of knowledge). Harmonic grammar (Legendre et al., 1990) posited that perceived acceptability could be idealised as levels of neural activation associated with competitor utterances (based on the computational notion of neural nets, rather than a biological model). A successor theory, Optimality Theory (see e.g. Prince and Smolensky, 1997) formalised the selection process as a

[7] *The Madisonian* Newspaper, Montana, February 1889. [http://www.stcboston.org/archives/articles/nouns.shtml]

[8] As discussed on page 12 of section 1.3, the relationship between perceived acceptability and frequency of use is far from straightforward.

strict hierarchy of discrete constraints, driven by a tension between brevity and precision of expression (termed economy and faithfulness, respectively). Vigliocco and Hartsuiker (2002) too speak of a common assumption in processing models of the twin objectives of economy and accuracy.

Dowty's model of argument selection is also based on competition: entities compete with each other for grammatical positions, with their goodness of fit being expressed in terms of their similarity to proto-roles. For example, in (4.7) Questionee[9] is variously realised as a direct or indirect object, despite identical semantics. As an utterance is formed, the entities that will become arguments are in competition for the central syntactic positions of subject, direct object, and indirect object. In (4.7a), the absence of a more likely candidate for direct object allows the Experiencer to jump into this position (the Question role is already incorporated into the semantics of the verb).

(4.7) a. On the ferry coming home again a fortnight later, **some French boys questioned her**$_{\text{dobj}}$ about what had happened to her hair.[10]

b. On the ferry coming home again a fortnight later, **some French boys asked her**$_{\text{iobj}}$ **questions**$_{\text{dobj}}$ about what had happened to her hair.

Such competition may also be seen to operate at the level of the clause (consider the alternatives in 4.2). In strongly transitive events, there is a strong asymmetry between the nature of participation of the causing and affected entities. More often than not, Agents are also human (high on a scale of natural saliency), and so strive to appear at the head of a sentence. So in such situations, a single argument structure is usually realised unchallenged as a sentence in which the Agent is the subject and the Patient is a direct object. Competition is more apparent when considerations of involvement and discourse are in conflict, or when one or the other is neutralised (i.e. all participants have a similar degree of involvement; or of saliency). In such cases, more than one structure is often attested within a single language, or between languages that otherwise have similar transitive structures.

Verbs of transfer of possession are a good example of this – consider *give/receive*. As two human participants are typically sentient volitional causers, and so have a similar claim on the ProtoAgent, the choice of verb allows either a Recipient (*buy/receive*) or Donor subject (*sell/give*). Commercial transaction verbs are more finely balanced since the seller, while being the Donor of the more prominent goods, is typically also the recipient of payment. There is also competition between the Recipient (stronger in terms of saliency) and Patient (stronger in terms of participation) – hence the dative diathesis

In situations such as *resemble* or *meet*, there can be perfect symmetry between the participants on the nature of their involvement, and subtle differences in meaning can affect which argument structure wins out. For example in (4.8) the two variants may seem interchangeable. Assuming the meeting in question was not planned, neither participant is more causing or affected than the other. With involvement neutralised as a factor, saliency and discourse considerations may drive the choice of one structure over another. In (4.8b) *Mary* is more conspicuous. This may be because the speaker identifies more closely with her (she is more salient), or because she was mentioned more recently in the preceding context (she provides coherence to the discourse). Of course an alternative interpretation of the (4.8b) variant is that the nature of involvement does differ between the two participants – for example Mary may have organised the meeting, or travelled to meet Paul at his home.

[9] In this discussion terms such as Questionee are used as labels of convenience for typical participants in an event, without making any theoretical claims on their status.

[10] BNC text B1Y: Biography, Molly Horne, 1989, *Life, love and laughter*.

(4.8) a. Mary and Paul$_{subj}$ met (each-other)

b. Mary$_{subj}$ met (with$_{oblq}$) Paul$_{dobj}$

An almost identical pattern or relatively free variation can be seen in four other languages: German (4.9), Spanish (4.10), Chinese (4.11) and Irish (4.12). But in none of these languages is it possible to conjoin strongly asymmetric arguments – e.g. in English a form such as *Tom stole a car* cannot alternate with **Tom and a car stole.*

(4.9) a. Maria und Paul$_{subj}$ traffen (sich)
Mary and Paul met REFL

b. Maria$_{subj}$ traff (mit$_{oblq}$) Paul$_{dobj}$
Mary met (with) Paul

(4.10) a. Maria y Paulo$_{subj}$ se han encontrado
Mary and Paul REFL have met

b. Maria$_{subj}$ (se) ha encontrado con/a Paulo$_{oblq}$
Mary REFL has met with/to Paul

(4.11) a. Mali Baoluo$_{subj}$ jianmian le
Mary Paul meet PFV
玛丽保罗见面了

b. Mali$_{subj}$ gen Baoluo$_{oblq}$ jianmian le
Mary with Paul meet PFV
玛丽跟保罗见面了

(4.12) a. Bhuail Máire agus Pól$_{subj}$ le chéile
met Mary and Paul together

b. Bhuail Máire$_{subj}$ le Phól$_{oblq}$
met Mary with Paul

Verbs of perception, emotional response or opinion are also moderately asymmetric predicates, since neither Experiencer nor Stimulus are strongly Agent-like or Patient-like. An Experiencer can be viewed as dynamic in some way, since there is a certain cognitive activity involved; but conversely it can be seen as slightly affected, as a change in cognitive state has been brought about by the Stimulus. In this last sense the Stimulus is a cause, albeit an inactive one; but it can also be seen as affected in the sense that knowledge about it has changed. For verbs of perception, many languages have pairs that realise Experiencers as subjects and Stimuli as objects (due to the stronger identification that speakers feel with Experiencers, which are by definition animate, and so relatively high on an anthropocentric saliency scale), but express the difference between active and inadvertent perception: compare *look, ansehen* (German)*, mirar* (Spanish)*, kan* 看 (Chinese), *féach* (Irish); and *see, sehen, ver, kanjian* 看见, *feic*. In cases where the Stimulus is unusually active, the linking of roles can be reversed, e.g. *Margaret Thatcher appeared to Christy Moore in a dream*.

The realisation of situations of liking and disliking is particularly various, both within and across languages. While many languages have verbs that express these relations from both points of view (e.g. *like/loath* compared to *please/disgust*), there is often a strong bias one way or the other.[11]

[11]For example, in the British National Corpus, the verb token *like* outnumbers *please* by about 20 to 1. In the German COSMAS corpus, *gefallen* 'please' outnumbers *mögen* 'like' in written text. A Chinese Google search (using the simplified orthography) found *xihuan* 喜欢 'like' to be 125 times more common than *quyue* 取悦 'please'. In the Corpus del Español, *desagradar* 'displease' is outnumbered 25 times by *odiar* 'dislike', in 20th century texts.

To take verbs of liking as an example (4.13), the Experiencer is commonly realised as a subject in English and Chinese, and as an indirect object in Irish, German and Spanish.[12] The Stimulus is a direct object in English and Chinese, and subject in Irish, German and Spanish.[13]

(4.13) a. Dustin likes cream cake

b. Dasiting$_{subj}$ xihuan naiyou dangao$_{dobj}$
Dustin like cream cake
达斯汀喜欢奶油蛋糕

c. Is maith$_{pred}$ le Dustin$_{iobj}$ cáca uachtair$_{subj}$
be good with Dustin cake cream

d. Sahnekuchen$_{subj}$ gefällt dem Dustin$_{iobj}$
cream-cake please to-Dustin

e. A Dustin$_{iobj}$ le$_{iobj}$ gusta tarta de nata$_{subj}$
to Dustin to-him please cake of cream

As was discussed in the introduction, and in section 2.5, *like* is much more common in English than *please*, and is often said to differ semantically (*like* is characterised as a state, and *please* as an activity). However, in some cases the two verbs can be semantically interchangeable, and the choice driven by factors of pragmatics. A stative use of *please* may be chosen to allow the Stimulus to appear at the head of the sentence (again, for reasons of coherence or saliency), over the otherwise more faithful *like* (e.g. 4.14, repeated from page 48).

(4.14) a. Multinationals have in the past been checked by laws which require Indians to take majority shareholdings in foreign companies operating in India. These regulations have, for instance, forced IBM out of India altogether. But the latest scheme to boost exports and control imports does not **please** India's software industry.[14]

b. ... India's software industry does not **like** the latest scheme to boost exports and control imports ...

Detransitive diatheses (also known as unaccusative, ergative, anti-causative and inchoative constructions, Levin, 1993, p.30) are another neat illustration of such a competition, and the differing ways in which languages resolve it. They are used when a single Patient-like entity is the only involved participant in the situation to be described, because an Agent-like participant is either absent (e.g. *a star burned out* in 4.15 below), or is omitted because it is unknown, unimportant, or understood (e.g. *it broke*).

(4.15) In the more likely case that [intelligence] typically evolves more slowly than stars, it would not have time to develop before **a star burned out**, Homo sapiens being a lucky exception.[15]

[12]There are three common phrases used in German to express liking, depending on the nature of the Stimulus: *gefallen* 'please' generally for objects, *gern machen* 'enjoy doing' for actions, and *mögen* 'like' for a restricted class of objects – see Wolf (1960).

[13]The clausal analysis of the Irish copula *is* remains controversial (Dónall Ó Baoill, personal communication), variously being described as Verb-Predicate-Subject (Genee, 1998, p.23), Verb-Predicate-Object (Noonan, 2004, p.225) or Particle-Predicate-Subject (Carnie, 1995, p.26) (for Carnie, *is* is "a complementizer particle and not a verb").

[14]BNC text B7H: *New Scientist* magazine, 1991.

[15]"Astronomer Puts Better Odds on ETs", *ScienceNOW* magazine, 31st August 1998.

(4.16) a. Miss, we haven't even got our protractor yet. Wait a sec. I did have a erm, protractor, **but it broke**.[16]

b. ..., but someone broke it.

c. *..., but it wrecked/knackered.

d. ..., but it got wrecked/knackered.

According to considerations of saliency and discourse coherence, the Patient-like entity should appear at the head of the sentence (a position occupied by subjects in most languages), since there are no other arguments competing with it – by default it is the most prominent. But according to the nature of its participation it will tend to be realised as a direct object (whose canonical position follows the subject in most languages). In English, a subject is obligatory in every declarative sentence. Passivisation allows canonical objects to appear in the subject position, but the construction is longer than an active sentence, and so due to a general preference for economy, may be dispreferred. For common verbs such as *break*, detransitive diatheses have developed, but these are lexically limited – if the verb in (4.16a) is replaced with *wreck* or the more colloquial *knacker*, the lack of availability of the detransitive diathesis (4.16d) forces a passive (4.16e) construction, if an Agent is not to appear.[17]

German differs from English in allowing free fronting of objects and other complements, but in common with English, has a structural requirement for a subject. There are some detransitive verbs such as *brechen*, but they are rarer than in English (Durrell, 2002, p.376), and many causative/detransitive verb pairs exist also, such as *kaputt machen* 'break' and *kaputt gehen* 'get broken'.

(4.17) a. Mein Winkelmesser$_{subj}$ ist kaputt gegangen
my protractor is broken gone
'My protractor broke'

b. Meine Freunde$_{subj}$ haben meinen Winkelmesser$_{dobj}$ kaput gemacht
my friends have my protractor broken made
'My friends broke my protractor'

In a structurally flexible language like Chinese, matters are much simpler. By the general processes of topic fronting applied to an object (where an argument appears before the entire verbal phrase, without changing in its grammatical function), and argument dropping applied to a subject (see page 81), an equivalent sentence to (4.16a) can be produced (4.18a).

(4.18) a. wo de liangjiaoqi$_{dobj,topic}$ pohuai le
me ATTR protractor break PFV
'My protractor broke'
我的量角器破坏了

b. you ren / wo / ta$_{subj}$ pohuai le wo de liangjiaoqi$_{dobj}$
there-be person / me / s/he break PFV me ATTR protractor
'Someone/I/he/she broke my protractor'
有人／我／他破坏了我的量角器

[16] BNC text KNY: Recorded secondary school classroom conversation.

[17] An interesting question that remains open, is why the detransitive diathesis does not apply to all transitive verbs in English. This selectivity appears to be quite arbitrary, even with semantically homogenous classes of verbs such as those of breaking – for example *smash* allows the detransitive diathesis, while *crush* does not. Since English lacks case marking, a sentence using a detransitive verb can result in temporary ambiguity. For example in an utterance beginning *The car smashed* ... it is not yet clear whether the car is an Agent or a Patient. It may be that English has retained detransitive/non-detransitive pairs such as *smash*/*crush*, so that speakers have the means to avoid this ambiguity when necessary.

Spanish has a general detransitive form (see Zagona, 2002, p.40) by which a canonical object appears in the subject position and its only marking is the *se* reflexive pronoun (so it is a more economical choice than a passive). The evidence that the *transportador* is a syntactic subject, rather than a fronted object, is the singular agreement of the verb *rompió* (3rd person singular) seen in (4.19a), rather than *rompieron* (3rd person plural).

(4.19) a. Mi transportador$_{subj}$ se rompió
my protractor REFL broke
'My protractor broke'

b. Mis amigos$_{subj}$ rompieron mi transportador$_{dobj}$
my friends broke my protractor
'My friends broke my protractor'

As these examples show, many languages broadly agree on how to present situations in which there is a clear asymmetry between causer and caused, where participation is the predominant factor. The most significant point of divergence in argument structure, both between and within languages, is in situations in which there is only moderate asymmetry between arguments.

4.3 A Numerical Model of Linking in English

As just discussed, there is varied evidence that language processing in general and that production of argument structures in particular are neither strictly serial nor modular. Here a proof-of-concept model is presented, that takes Dowty (1991) as its starting point. In this section the proposed model is first situated relative to existing theories of constraint satisfaction and acceptability in language. The points on which the model differs from Dowty's are discussed and defended. This model characterises proto-roles in terms of notions of involvement, as Dowty does, but supplements this with a set of proto-positions that relate pragmatic and discourse structuring factors to serial ordering. Individual factors are derived from underlying speaker construals of the event of interest, rather than being argument properties entailed by the verb in question. And argument prototypicality is aggregated over the event to yield a measure of clause prototypicality, which is then interpreted as a measure of wellformedness.

In section 4.3.1 some limitations of Dowty's model discussed in chapter 2 are treated at greater length, and it is argued that a three-way partition of proto-roles better reflects the behaviour of nuclear grammatical relations in English. This model also applies differential weights to each determining factor (section 4.3.2), which are derived from correspondences seen in the experimental materials used in the annotation exercise (see section 3.4). The operationalisation of the model is described in section 4.3.3, and it is evaluated against predictions from the literature and from acceptability judgements in section 4.3.4.

The model presented here makes several simplifying assumptions about the nature of the interaction between the satisfaction of individual constraints. The majority of features are treated as binary quantities, though graded values may also be appropriate. For example, some living things may be conceptualised as more "animate" than others (see page 84 of section 3.2.1.5). Affectedness might also be considered a graded quantity: entities that are brought into and out of being in the course of an event (i.e. Product and Resource roles, or Factitive in Fillmore's terminology) are the most entirely affected; while less dramatic physical change may constitute a lower degree of affectedness, whether extrinsic (e.g. movement) or intrinsic (e.g. damage or repair).

The effects of contributing factors are treated as though they were independent, though this is not the case (see the confusion matrices 3.4 and 3.7 in sections 3.2.2.4 and 3.2.3.4 for evidence of confounding in the English materials). The effects of these factors are combined linearly, even though other functions (logical or numerical) are possible.

Without such assumptions, the space of possible numerical or algorithmic instantiations of models is large. Classical generative models assume that the realisation of argument structure is a strongly constrained process internal to syntax. Multiple syntactically wellformed candidates are generated on the basis of selected lexemes, and then semantically inconsistent ones (e.g. **himself$_i$ washed him$_i$* or **Golf played John*) are rejected at the interface with a semantic subsystem (Chomsky, 1986, 1995). Computationally, this is equivalent to generation using regular expressions, which are then constrained by further logical conditions.

Optimality theory (OT) assumes a strict hierarchy of constraint violations. This is algorithmically equivalent to a binary decision-tree which yields a single acceptable ("optimal") structure, discarding all other candidates as unacceptable ("suboptimal"). Variations on OT have also been proposed. Stochastic Optimality Theory (see e.g. Boersma, 2004) retains strict ordering of constraints, but introduces noise into the process of satisfaction to account for random variation. Linear Optimality Theory (Keller, 2006) gives constraints weights (rather than ranks) that represent their relative severity, which in turn allow individual effects to be combined numerically. The aggregate level of violation is then interpreted as a measure of acceptability. Harmonic theory (Legendre et al., 1990) is a predecessor theory to optimality theory, that is based on neural nets. It uses levels of activation in neural nets to represent the *harmony* (suitability or acceptability) of a structure. The relative weights of constraints and their interactions are encoded in the form of weights on connections between nodes in the network.[18]

Dowty's model of argument realisation is based on the operation of counting the number of properties (entailments) satisfied to evaluate similarity to a prototype, and using logical rules to select among candidates. This has similarities to the notions of distance, similarity and class membership used in machine learning techniques such as clustering and case-based reasoning.

Combinations of these models are also conceivable. It may be that certain categorial constraints operate strictly (in the manner of OT), to rule out some candidate structures, but that selection among the remaining structures allows interaction between constraints. Constraint satisfaction also need not be based on linear functions, or even be deterministic. For example, optimisation may operate in the manner of techniques like simulated annealing, which find local minima in functions, yielding "near-optimal" solutions.

The model presented here takes Dowty's model of argument selection as its theoretical starting point, but is more similar to Keller's Linear Optimality Theory in its operationalisation. It differs from Dowty's model in several other important respects. Firstly, Dowty argues that the characteristic properties of roles should be defined in terms of entailments of verb argument positions. If this is interpreted strictly, it can lead to unwanted results, such as that the subject of *murder* does not have the

[18]Neural nets are a general-purpose machine learning methodology, and so can be expected to solve a wide range of constraint satisfaction problems. Typical implementations are of dramatically smaller scale than networks in the brain. For example, a successful neural net model of property-based conceptual categorisation by McClelland and Rogers (2003) requires about two to three nodes per concept in the mental lexicon. Only a few hundred thousand such nodes would be required to represent the semantics of a typical lexicon of 50,000 to 100,000 lexemes (see e.g. Levelt, 2001). While the equivalent neural structure has not been conclusively identified, areas of the brain that are known to be particularly involved in language are much larger. Broca's area (which is required for language production) can contain over 250 million neurons in adults (Uylings et al., 2005, p.721). The fusiform gyrus (which has a function in conceptual categorisation) may consist of 150 million neurons in healthy adults (assuming a similar neural density, and volumes cited by Onitsuka et al., 2003). The fact that a theory can be implemented using neural networks does not in itself confer biological plausibility.

volitional feature of ProtoAgent (even though this is typically the case) since someone can be forced to commit a killing (and so be intentionally but not volitionally involved). A possible modification to his scheme would be to count the default, or typical entailments of a verb position. However, in section 1.1 it was suggested that grammatical processes may be governed by the properties of the specific situation being described, rather than by the default lexical semantics of the verb used to express it. Experiments by Bornkessel and Schlesewsky (2006) also indicate that the semantics of argument structures are processed by listeners *before* the verb is encountered (the experiments use German, which has verb-final structures and case-marking of noun phrases, allowing listeners to construct underspecified event descriptions before the verb is reached). In the model presented here, the characteristic features are instead stated in terms of properties of the event and its participants. This takes a different perspective on sentence production, in which the speaker's construal of the event selects the verb, rather than the verb itself imposing a construal.[19] In this respect the model is opposed to the view of Levin and Rappaport Hovav (2005, p.237) that such characteristic features should be "event-based notions ... [that a] verb imposes on its arguments by virtue of the parts they play in the event it describes" and that "lexical semantics of the root determine ... argument realisation" (*ibid* p.241). This construal-centric approach can also be considered more informative, since it describes properties of the specific situation instance in question, rather than default properties of that class of situation, or redundant facts about its participants (e.g. that one of them is animate).

Further, rather than simply counting the number of atomic features that apply to a participant (in this Dowty follows Rosch, 1973), each property is given a characteristic weight to represent its effect magnitude. These parameters are derived from measures of association between individual features of arguments, and the semantic role labels with which these arguments were annotated in the exercise in section 3.4 (the precise manner of their derivation is described in section 4.3.2). Furthermore, in some cases the feature values themselves are graded. For example the Affect feature will have one of three values (positive, neutral and negative) and points on the saliency hierarchy will also be ordered.

Dowty argued that while discourse related notions do affect argument realisation, semantic roles is not the correct rubric for them. However, as was discussed in section 2.6, some of his proto-entailments seem to stray beyond the description of the nature of a participant's involvement (the movement entailments have no distinguishing function in the absence of a Ground; sentience, in itself, is not an attribute of participation, but rather of a participant). While discourse features are clearly of a distinct sort from involvement features, this proposal takes the view that an adequate account of English argument structure cannot be made without appealing to elements of pragmatics.

Hence two parallel proto-systems operate: a set of proto-roles (that correspond to canonical grammatical relations – most strongly determined by considerations of involvement); and a set of proto-positions (that correspond to serial order – more strongly determined by pragmatic considerations). The set of proto-positions will be defined in terms of properties of anthropocentric saliency[20], of information structure (the given/new distinction – see Prince, 1981), and of argument heaviness.

This characterisation of proto-positions ignores one important notion that plays a considerable role in the realisation of argument ordering. Rhetorical focus was not manipulated in any of the experiments performed here, and so there is no principled basis on which to include it in this parameterised model.[21] A complete model would have to account for this effect.

[19] Of course, during comprehension the verb does play a large part in imposing a particular construal, but the experiments of Bornkessel and Schlesewsky (2006) suggest that default construals of grammatical positions can operate independently.

[20] This notion of speaker-centric saliency might also be termed speaker-identification or speaker-empathy (see page 58 for a discussion of such ideas from Fillmore, Kuno and Croft).

[21] Rhetorical focus was not investigated for two reasons. One is that there still remains a great deal of disagreement, not just

The idea that a clause which had strongly proto-typical arguments would itself be a strongly proto-typical sentence was implicit in Dowty's 1991 model. In the model presented here, it is made explicit. The prototypicality of a candidate clause structure is the average of the prototypicality of each of its participants, relative to the grammatical relations and serial positions they hold. And the prototypicality of a clause is a measure of its appropriateness to describe the situation in question (this has obvious parallels to the idea of harmony just discussed). One consequence of this is that the relative prototypicality of complete candidate structures can be compared, and the most acceptable expression identified. A further consequence is that candidate clauses based around different verbs can also be compared.

Finally, Dowty's model, in common with others (e.g. van Valin and LaPolla, 1997), proposes two proto-roles that correspond most closely to the English grammatical relations of subject and direct object. In this model three roles are used. A ProtoExperiencer is posited to correspond to the indirect object, and so supply an adequate analysis of structures such as the dative diathesis, and verbs of communication, that lay beyond the remit of the original theory of proto-roles. The rationale for using three roles, and for its characterisation as a ProtoExperiencer, are detailed in section 4.3.1.

It must be stressed that this model is not entirely unique. A system with three proto-roles has been suggested by Primus (1999). Various authors have integrated semantic and pragmatic factors into accounts of argument structure, including heaviness, information structure and salience hierarchies (see e.g. Bresnan et al., 2001; Keller and Asudeh, 2001; Bresnan and Nikitina, 2003; Wunderlich, 2006). Numerically based models have been built that base their parameters on empirical studies (e.g. Bresnan et al., 2007), and that interpret numerical outputs as measures of appropriateness or acceptability (Legendre et al., 1990; Keller, 2006).

The model presented here does not account for the mechanism of choosing when to say what – something that is primarily an executive function. Nor does it address the means by which larger episodes are segmented into units that correspond to spoken sentences.[22] It assumes that the meaning of a longer discourse has already been segmented into individual "messages" that correspond in extent to a spoken sentence (Bock and Levelt, 1994, pp.945-947), and that the speaker has already chosen what aspects of the situation are to be represented as entities (and so participants) and what the event of interest is (what becomes the predicate). This process of "subchunking" (Croft, 2007) is a complex one that is effected by diverse factors including the nature of reality, the speaker's goals and construal, and cultural conventions (*ibid.* sec. 2.2).

Neither does the model give a complete account of English argument structure realisation. The identification and relative ordering of oblique arguments is not provided for. The selection of verb and preposition lexemes is not addressed, and the models measure of acceptability does not take the effects of lexeme frequency and brevity of expression into account. However, the author is unaware of any such work that attempts to provide a general account of English argument structure realisation including all the nuclear grammatical relations, and that has the potential to explain non-canonical orderings such as the passive.

The purpose of this model is rather as a proof of concept. It uses a rudimentary linear numerical model, assumes independence between determining factors, and derives its parameters from an

on terminology (consider topic/comment; topic/focus; ground/focus; theme/rheme), but on the proper analysis of emphasis relations within sentences (see e.g. Kruijff-Korbayová and Steedman, 2003). Secondly, intonation is one of the principle mechanisms for signalling focus relations (see e.g. Steedman, 2007), so a proper investigation of this factor would have required spoken experimental materials to be gathered, and would have presented difficulties for an experiment administered to large numbers of participants over the web.

[22] This is known as "text planning" in the natural language generation literature. For example Rhetorical Structure Theory (Mann and Thompson, 1988) is used widely for the task (Piwek and van Deemter, 2006).

extremely small corpus of active English sentences. But despite these simplifications, it succeeds in accounting for many of the English argument structure phenomena that have been raised in the literature. The objective here is not to develop a formalism that necessarily duplicates the processes and structures of linguistic cognition.[23] Rather the model should simulate the choices made by speakers when converting a thought into an utterance, to answer the question: once a person has decided what to say, how do they decide how to say it?

4.3.1 Number of Proto-Roles

Dowty's model based around the generalised semantic roles of ProtoAgent and ProtoPatient cannot account for some common English sentences structures, especially those that involve animate participants whose involvement is intermediate between these two poles. The properties of each proto-role (modified from Dowty, 1991, p.547) are listed below, with the condition of independence of existence omitted (Dowty expressed doubt about these entailments, and in the discussion here, they are not a differentiating factor, as none of the participating entities are brought into or out of being during the situations described – see page 52). As discussed in section 2.6, the movement entailment is formulated in such a way that it has no differentiating function. It can never be a distinctive feature, since if any participant is moving relative to any other participant, then *all* participants are moving relative to some participant. For that reason, it is interpreted here as relative to a ground, in the sense of Talmy (1985).

(4.20) Contributing properties for the Agent Proto-Role:

a. volitional involvement in the event or state

b. sentience (and/or perception)

c. causing an event or change of state in another participant

d. movement (relative to ground)

(4.21) Contributing properties for the Patient Proto-Role:

a. undergoes change of state

b. incremental theme

c. causally affected by another participant

d. stationary (relative to ground)

Dowty's principles (1991, p.576) are also repeated here, with the non-discreteness condition omitted as it is does not affect argument selection (see 2.176d, page 53 for the non-discreteness condition):

[23] In the absence of better understanding of cognitive representations, a formalism is more likely to conform to notational and theoretical convenience than to cognitive reality (though of course psychological reality is not the only valid objective of semantic or other formalisms – see Partee, 1979). While formal models may reflect logical thought and reasoning, and language is obviously symbolic, much of our cognitive activity is very unlike this – for example as Gärdenfors (2000, p.236) comments, it is unlikely hand-eye coordination is governed by "solving symbolic equations". No commitment need be made on the form of conceptual representations here, as whether internal conceptual representations are visual, verbal, symbolic or a combination of these, it is an empirical fact that people can translate between these representations (e.g. we can hear a spoken word, and then visualise it or retrieve its written form).

(4.22) a. Argument Selection Principle: In predicates with grammatical subject and object, the argument for which the predicate entails the greatest number of ProtoAgent properties will be lexicalized as the subject of the predicate; the argument having the greatest number of ProtoPatient entailments will be lexicalized as the direct object.

b. Corollary 1: If two arguments of a relation have (approximately) equal numbers of entailed ProtoAgent and ProtoPatient properties, then either or both may be lexicalized as the subject (and similarly for objects).

c. Corollary 2: With a three-place predicate, the nonsubject argument having the greater number of entailed ProtoPatient properties will be lexicalized as the direct object and the nonsubject argument having fewer entailed ProtoPatient properties will be lexicalized as an oblique or prepositional object (and if two nonsubject arguments have approximately equal numbers of entailed P-Patient properties, either or both may be lexicalized as direct object).

As was discussed in section 2.5 the distribution of indirect objects cannot be accounted for adequately by these principles. Here it will be treated in more detail, with reference to typical ditransitive situations of transfer of possession and of communication.

In a typical event of agreed transfer of property, the Donor only has one ProtoPatient property (being stationary) and is the best candidate for ProtoAgent with three properties: being volitionally involved, sentient and causing. The Gift has the single Agent-like property of moving, but is the optimal Patient with two properties: being causally affected and undergoing a change of state (of position and possession). The Recipient has only one ProtoPatient property (being stationary) and two ProtoPatient properties (volitional involvement and sentience).[24] Interpreting Dowty's selection principle (4.22a) strictly, the only possible realisation of such an event is (4.23). Since the nonsubject participant *Cecilia* has fewer ProtoPatient features than the other non-subject (*£30,000*), it must be realised as a prepositional phrase (4.22c).

(4.23) Bertie$_{\text{Donor}}$ gave £30,000$_{\text{Gift}}$ to Cecilia$_{\text{Recipient}}$

On a more permissive interpretation of Dowty, the Recipient and Gift can be considered to have "approximately equal numbers" of ProtoPatient properties (following 4.22c), and two further configurations are licenced: the wellformed (4.24) and the illformed (4.25).[25]

(4.24) Bertie$_{\text{Donor}}$ gave Cecilia$_{\text{Recipient}}$ £30,000$_{\text{Gift}}$

(4.25) #Bertie$_{\text{Donor}}$ gave Cecilia$_{\text{Recipient}}$ PREP £30,000$_{\text{Gift}}$

On this permissive reading of Dowty, both Recipient and Gift can also be realised as direct objects. The configuration in (4.26) is then not ruled out, and the possibility of conjoining such pairs also admits (4.27-4.28).

(4.26) #Bertie$_{\text{Donor}}$ gave £30,000$_{\text{Gift}}$ Cecilia$_{\text{Recipient}}$

[24] Typically the Recipient in such a situation is aware, and agrees to participate, but involuntary instances of receipt are possible (e.g. one can make a gift to someone in a coma, or to an organisation). Interpreting Dowty's properties strictly – that is, entailments that invariably hold of an argument – the Recipient would thus have no ProtoPatient properties. Similarly, the verb *give* does not require that its Donor subject and indirect object Recipient are stationary, though that is a typical construal.

[25] The unspecified preposition is notated as PREP.

(4.27) #Bertie$_{\text{Donor}}$ gave Cecilia$_{\text{Recipient}}$ and £30,000$_{\text{Gift}}$

(4.28) #Bertie$_{\text{Donor}}$ gave £30,000$_{\text{Gift}}$ and Cecilia$_{\text{Recipient}}$

Finally, Dowty does not make constraints on ordering explicit, so (4.29-4.30) might also be considered valid constructions in his model.

(4.29) ?Bertie$_{\text{Donor}}$ gave to Cecilia$_{\text{Recipient}}$ £30,000$_{\text{Gift}}$

(4.30) #Bertie$_{\text{Donor}}$ gave PREP £30,000$_{\text{Gift}}$ Cecilia$_{\text{Recipient}}$

Some of these unlicensed configurations can be addressed with rather minor changes to Dowty's proposal. A specification that conjoined arguments must be similar to *each other*, as will be proposed in the model presented here, could rule out sentences such as (4.27-4.28), and one that prepositional phrases must follow bare noun phrases would exclude (4.29-4.30).

However, as these examples show, the Recipient argument, which in traditional grammars would be associated with the indirect object, has a distinctive distribution. Treating it as any other prepositional or oblique phrase will rule out the wellformed (4.24), and treating it as a marginal direct object (as Dowty does) admits the illformed (4.25) and (4.26). The benefactive diathesis, which involves an Agent, Beneficiary and some sort of Theme or Patient is subject to the same problems.

Similarly, Dowty's account provides false predictions for the argument structures realised for some communicative events. The Message and Addressee arguments of a verb such as *tell* pattern in the same way as dative and benefactive arguments (i.e. *tell someone your story* and *tell your story to someone* are both possible). The Speaker is the ProtoAgent (volitional; sentient; causing), Message is the ProtoPatient (incremental theme; causally affected) and the Addressee is a lesser Agent (sentient; and often volitional) and lesser Patient (no properties). Thus, even with the minor modifications just mentioned (mutual similarity as a constraint on conjoining; enforcing NounPhrase–PrepositionalPhrase ordering) it would be predicted to be realised in some of the same illformed constructions as seen above with the dative example:

(4.31) Albert$_{\text{Speaker}}$ told the story$_{\text{Message}}$ to Charlie$_{\text{Addressee}}$

(4.32) Albert$_{\text{Speaker}}$ told Charlie$_{\text{Addressee}}$ the story$_{\text{Message}}$

(4.33) #Albert$_{\text{Speaker}}$ told Charlie$_{\text{Addressee}}$ PREP the story$_{\text{Message}}$

(4.34) #Albert$_{\text{Speaker}}$ told the story$_{\text{Message}}$ Charlie$_{\text{Addressee}}$

For a verb such as *talk*, which does not take a Message argument, the Speaker is again the ProtoAgent, and is realised as the subject. Dowty does not specify a minimum number of proto-properties, so one assumes that Addressee is then left to be the ProtoPatient, despite having none of its entailments. This would yield the unlicensed structure (4.35). If both participants are similarly active in the conversation, the conjoined variant (4.36) could result, but no possible construal of a talking situation can result in the common construction (4.37) in which the Addressee is expressed as a prepositional phrase.

(4.35) #Enda$_{\text{Speaker}}$ talked Eamonn$_{\text{Addressee}}$

(4.36) Enda$_{\text{Speaker}}$ and Eamonn$_{\text{Speaker}}$ talked

(4.37) Enda$_{\text{Speaker}}$ talked to Eamonn$_{\text{Addressee}}$

The solution to this issue proposed here is to add a third proto-role to correspond to the distributionally distinct grammatical function of indirect object in English.[26] As was discussed in sections 2.6 and 3.4, a similar proposal is made by Primus (1999; 2002). She characterises the Proto-Recipient as an intermediate category that is conceptually dependent on the ProtoAgent and ProtoPatient (the argument being that to have a Recipient, both an Agent and a Theme are presupposed), and that is most closely identified with the dative case in German. Primus' Proto-Recipient encompasses the traditional roles of Recipient and Beneficiary, but still applies to quite a restricted class of arguments (for example it might not include the Addressee above).

To account for the similar syntactic realisations given to Recipients, Beneficiaries and Experiencers in English, the broader ProtoExperiencer is proposed here in this model. It also differs from Primus in being characterised in a manner more similar to Dowty's, defined in terms of prototypical features (that are introduced in the following section). As the projected figures for inter-annotator agreement on proto-role inventories suggested, an annotation exercise using a Agent-Patient-Experiencer schema would be expected to result in considerably higher levels of normalised agreement (relative to expected agreement, for that set of categories) than the traditional role inventory (see table 3.30, page 142). It would also give similarly high levels of agreement to that of Dowty's original Agent-Patient inventory and Primus' Agent-Patient-Recipient, but with wider coverage of English argument structures.

Finally, one criticism of Dowty, that can also be levelled at this account, is that the proto-roles are really characterisations of grammatical roles under a different name. ProtoAgent might simply be called a semantic characterisation of subject, and this is also suggested by the fact that Dowty's account (1991) used Keenan's universal definition of subject (1976) as a one of its starting points. However there are two reasons for separating the notions of semantic and grammatical roles. Firstly, this account characterises the proto-roles purely in terms of participation, while the grammatical roles also have discourse functions (here handled by the proto-positions). Secondly the two notions must be kept separate if the lexicon can store the argument structure of verbs with non-canonical argument structures – for example the verb *receive* must encode the fact that its subject is associated with a ProtoExperiencer, and *undergo* must specify that it takes a ProtoPatient-like subject.

4.3.2 Parameterisation

There any many ways in which the parameters of a model such as this could be derived. Considering just the empirical results available from this study, one possibility would have been to directly use the feature-coded materials and acceptability judgements from sections 3.2.2-3.2.5, to back-engineer a model of production for each of the constructions examined: the English dative/benefactive; the English and German passives; and the *bei* and *ba* constructions of Chinese.

[26] Agent, Patient and Experiencer-like participants receive prominent grammaticalisation in many languages (Palmer, 1994, pp.6-8). Spanish and German all have three categories of pronoun corresponding to the three grammatical functions of subject, direct object and indirect object. In Spanish these three functions are further distinguished: the subject is privileged in assimilating with the verb (by number and person agreement), while special forms of direct and indirect object pronouns (clitics) can appear immediately before the verb. English also has subject-verb agreement, while direct and indirect objects are the only categories that can appear without prepositional marking, and that can be fronted in a passive construction. The German cases of nominative, accusative and dative are based on these same categories, and the subcategorisation constraints of verbs and prepositions towards the noun phrases that they govern specify for these cases. Irish has a class of prepositional phrase constructions that are used to express Experience (e.g. *tá ocras orm* 'I'm hungry', literally *hunger is on me*). Given the genetic relationship between these languages, these regularities are perhaps not so surprising, but a similar pattern is seen in Chinese, a completely unrelated language which has no agreement or case features of any kind. The subject is positionally privileged in Chinese, having first claim to the preverb position. Like English, German and Spanish, an indirect object can jump into the postverb position (normally occupied by the direct object) and can appear without coverb (preposition) marking.

Table 4.1: Correlations and weightings for Proto-Roles

Role	**ProtoAgent**			**ProtoPatient**			**ProtoExperiencer**		
Dimension	***r***	***r*²**	***w***	***r***	***r*²**	***w***	***r***	***r*²**	***w***
Negative Outcome	-0.22	-0.05		**0.41**	**0.17**	**0.41**	-0.09	-0.01	
Positive Outcome	-0.18	-0.03		-0.16	-0.03		**0.45**	**0.20**	**0.25**
Cause	**0.79**	**0.63**	**0.55**	-0.50	-0.25		-0.13	-0.02	
Physically Affected	-0.14	-0.02		**0.49**	**0.24**	**0.59**	-0.19	-0.04	
Psychologically Affected	-0.25	-0.06		-0.04	0.00		**0.43**	**0.18**	**0.23**
Volitional	**0.59**	**0.35**	**0.31**	-0.51	-0.26		**0.21**	**0.05**	**0.06**
Aware	**0.40**	**0.16**	**0.14**	-0.32	-0.10		**0.35**	**0.12**	**0.15**
Possessor	-0.19	-0.04		-0.15	-0.02		**0.49**	**0.24**	**0.30**

r is linear coefficient of correlation, derived using the point-biserial measure
r^2 is the co-efficient of determination
w is weighting derived from the relative proportions of r^2

Instead, the feature-coded materials from the annotation study were used as a basis (81 active English sentences; see section 3.4). First of all, the annotations for the individual semantic roles used in the study were conflated into three proto-roles as follows (repeated from page 145):

(4.38)
- ProtoAgent: Agent and Instrument
- ProtoPatient: Patient, Percept and Theme
- ProtoExperiencer: Experiencer, Recipient and Beneficiary

Then measures of correlation were computed between each of these roles and each of the individual semantic features that the empirical studies suggested are associated with the expression of the nature of involvement (volition, awareness, causation, affectedness, positive/negative outcome, possession), and which have been suggested as drivers of argument structure in the literature. These three proto-roles are then related to the three principal grammatical positions of English: subject, direct object and indirect object.

Similarly, three serial positions (the argument before the verb; and the first and second arguments after the verb) were examined to see their association with pragmatic features: newness/givenness, anthropomorphic saliency and argument heaviness (i.e. length).[27]

This approach had several advantages. Firstly, it did not make use of the acceptability data, leaving it available as an independent data source for evaluation of the model. Secondly, since the input is a set of active sentences, the model can be expected to account for general sentence patterning in English, rather than just for a couple of specialised structures (the dative/benefactive and passive). Thirdly, the proto-roles and proto-positions provide a level of abstraction away from the very small quantity of material, providing the possibility of generalisation.

Given a larger set of materials a multiple linear regression could have been used to tease the independent effects of these features apart from their interactions (see e.g. Baayen et al., 2007). With the present set of materials, a naive assumption of independence is made, though many of the features are confounded (see tables 3.4 and 3.7 on pages 100 and 103).

The correlations of features of interest to each of the three proto-roles is given in table 4.1. All correlation coefficients *r* are linear coefficients, calculated using the point-biserial method (4.39).

[27] Probabilities might have been used in place of measures of correlations. However correlations have the advantage of providing a single measure that can capture both positive and negative effects of a feature.

Table 4.2: Correlations and weightings for Proto-Positions

Arg. Positions	**Preverb**			**First Postverb**			**Second Postverb**		
Features	***r***	***r*²**	***w***	***r***	***r*²**	***w***	***r***	***r*²**	***w***
Given	0.18	0.03	**0.24**	-0.15	0.02	**-0.21**	-0.10	0.01	**-0.50**
Saliency	0.30	0.09	**0.68**	0.28	0.08	**0.73**	0.10	0.01	**0.50**
Short	0.10	0.01	**0.08**	0.08	0.01	**0.06**	0.01	<0.01	**0.00**

r is linear coefficient of correlation, derived using the point-biserial measure
r^2 is the co-efficient of determination (sign is retained to signal the polarity of the effect)
w is weighting derived from the relative proportions of r^2

This is equivalent to taking the Pearson linear coefficient over two variables, one numerical and one binary.

(4.39) $r_{pb} = \frac{(\bar{x}_0 - \bar{x}_1)\sqrt{pq}}{s_x}$, where p and q are proportion of 0 and 1 values (so $p = 1 - q$). pq can also be expressed as $\sqrt{\frac{n_0 n_1}{n(n-1)}}$, $n = n_0 + n_1$

Here, the correlation coefficient is calculated over the individual clause arguments annotated in the exercise described in section 3.4, to determine the strength of association between a particular hand-coded feature, and the frequency with which arguments that bear that feature are annotated with a particular role label. Each feature is taken as the binary variable, and is correlated against the proportion of annotations that each argument received for the role in question.[28]

The square of the correlation measures, r^2, is the coefficient of determination, which is a measure of the proportion of mutual variance explained by a pair of variables. So, for example being a causal participant in an event or not accounted for 63% of the patterning in the annotation of ProtoAgent (i.e. by considering causation alone, it is possible to explain about two thirds of the decisions taken by annotators to use the role labels of Agent and Instrument).

In turn, weightings can be derived from the proportions expressed by coefficients of determination. For the sake of simplicity, only positive correlations are considered and an assumption is made that the effect of each feature is independent of all others. This last assumption does not hold for many of the features here (e.g. volition and awareness are expected to be confounded, and this was demonstrated in section 3.2.2.4), but the data quantities used here do not allow such interactions to be reliably quantified. The final weights (w) represent the proportion of variance that each feature explained, relative to the total variance explained. For example, assuming independence, a total 41% (0.17+0.24) of ProtoPatient annotation is explained by the features, Negative Outcome and Physically Affected, and these respectively constitute 41% and 59% of that total.

Correlations also established that each of these roles was most strongly associated with one of the three English grammatical relations: the mutual correlation between ProtoAgent annotations and subjecthood was 0.82; between ProtoPatient and the direct object was 0.79; and between ProtoExperiencer and the obliques was 0.43.

The corresponding correlations and weightings for the proto-positions are given in table 4.2. Since the number of determining features is smaller, both positive and negative contributions are considered. Here the correlations were derived directly from the hand-coding of the materials in the annotation exercise. Each feature is correlated against a binary variable for each position.

[28] Where necessary, non-binary variables are reduced to a binary distinction. For example the three-way positive/neutral/negative distinction is reduced to negative/non-negative and positive/non-positive.

For example, the measure of association between the feature Given and the Preverb position is that between two binary variables: given/new and preverb/postverb. In this model the taxonomy of Prince (1981) is followed, but alternative instantiations might allow for more than two values, by using a continuous measure of accessibility or topicality.

Argument heaviness is operationalised as Shortness, measured in terms of the inverse of the number of words (i.e. an argument of two words has a shortness value of 0.5; one of three words a value of 0.33). This is a rudimentary measure of heaviness, and one that takes phonological structure into account (e.g. by counting syllables) or measures syntactic structural complexity (e.g. by counting nodes in an argument's syntactic subtree) might be superior.

Anthropocentric saliency is operationalised with the following hierarchy (4.40). Numerical values are assigned to each of the categories by evenly dividing the space of numbers from zero to one.

(4.40) Speaker [0.93] > Hearer [0.79] > Other Humans [0.64] > Other Living [0.50] > Other Moving [0.36] > Other Concrete [0.21] > Abstract [0.07]

Again, many variations on such a hierarchy would be plausible. For example, a person known to the speaker or hearer might have higher natural saliency than strangers; and the saliency relation between speaker and hearer may change subject to politeness constraints (the speaker may choose to say *you and me* rather than *me and you*).

4.3.3 Operation of the Model

The typicality or suitability of a particular clause structure to describe an event is quantified as an aggregate of its participants' suitability to each argument structure position. Specifically, the suitability of a particular utterance u to describe an event e (4.41) is the arithmetic average over its n realised arguments of the suitability of each participant p to the grammatical position g in which it is realised.

(4.41) $$t(u,e) = \frac{\Sigma t(g,p)}{n}$$

The individual typicality or suitability of each participant is the arithmetic mean of suitability to its serial position s and to its canonical grammatical relation r (4.42). Currently no differential weighting is applied to the measures of serial position and canonical relation, though another version of the model could do so.

(4.42) $$t(g,p) = \frac{t(r,p)+t(s,p)}{2}$$

In absolute terms suitability is a measure of the similarity of the utterance to a prototypical transitive clause (for Croft, 1993, p.58, this is the "prototypical event", and it can also be compared the to the transitive construction from which other constructions are derived in construction grammar, Goldberg, 1995, p.117). In relative terms the utterance level measure is interpreted as a measure of its appropriateness as a description of the event in question, in a similar manner to Harmonic Grammar or Linear Optimality Theory.

The optimal utterance for a given event $o(e)$ can then be selected on the basis of this measure of suitability (4.43). The candidate utterances for an event *U(e)* are generated by taking all possible permutations of arguments with grammatical positions.

(4.43) $o(e) = argmax_{u \in U(e)} t(u,e)$

In its current form, the model assumes that all candidate verb lexemes are equally likely (e.g. *buy/peddle/sell/flog/pay/purchase*), and does not account for considerations of economy (a more complete model might reward brevity of expression and the use of familiar verbs over obscure ones).

The suitability of arguments to each of the three proto-roles (and so to canonical grammatical roles) are quantified using the parameters listed in table 4.1:

(4.44) $t(ProtoAgent, p) = 0.55(causal_p) + 0.31(volitional_p) + 0.14(aware_p)$

(4.45) $t(ProtoPatient, p) = 0.59(physicallyAffected_p) + 0.41(negativeOutcome_p)$

(4.46) $t(ProtoExperiencer, p) = 0.30(possessor_p) + 0.25(positiveOutcome_p)$
$+0.23(psychologicallyAffected_p) + 0.15(aware_p) + 0.06(volitional_p)$

Each of these functions returns a value between zero and one. For example, the typical causal participant of a *persuading* event, which is volitional, causal and aware, would have a suitability measure of 1 to the ProtoAgent. A participant who *falls asleep at the wheel*, would only have a ProtoAgent measure of 0.55, as the person in question would typically be causal, but neither volitional nor aware. It is the state of affairs at the conclusion of the event in question that forms the input to the model – so for a giving event the Possessor feature applies to the Recipient, and not to the Donor. Similarly, Affectedness may develop over the course of the event in question.

This instantiation of argument-level measures of suitability can be interpreted in terms of the *city block* or *Manhattan* distance commonly used in machine learning techniques to evaluate similarity or dissimilarity in geometric terms. They can also be compared to the conceptual spaces of Gärdenfors (2000), in which conceptual membership is quantified by the linear combination of values on discrete feature dimensions, and to many other models that quantify similarity of meaning in terms of geometric or graph spaces (e.g. Miller and Charles, 1991; Schütze, 1993; Lee Joon Ho et al., 1993; Landauer et al., 1998).

Similarly, the measures of suitability or typicality of each argument to serial positions can be quantified as follows. Negative correlations are handled by taking one minus the feature value, so that these functions also return a number between zero and one.

(4.47) $t(ProtoPreverb, p) = 0.24(given_p) + 0.68(saliency_p) + 0.08(short_p)$

(4.48) $t(ProtoFirstPostverb, p) = 0.21(1 - given_p) + 0.73(saliency_p) + 0.06(short_p)$

(4.49) $t(ProtoSecondPostverb, p) = 0.5(1 - given_p) + 0.5(saliency_p)$

Oblique phrases are treated differently, in that they are viewed as occupying a proto-position, but not occupying a proto-role.

Finally, coordination of arguments is also provided for, in the case that two participants are *equally or more* similar to each-other, than they are to their optimal argument realisation. For example, if two participants p and q are more typical of the ProtoAgent, than of either of the other two proto-roles, their mutual suitability (or similarity) relative to ProtoAgent can be calculated given as follows:

(4.50) $t(p, q, ProtoAgent) = 1 - |0.55(cause_p - cause_q)| - |0.31(volition_p - volition_q)|$
$-|0.14(aware_p - aware_q)|$

If this figure is equal or higher than each participant's individual suitability to the ProtoAgent role (4.44), they are candidates to be conjoined. The candidate utterance containing this conjoined argument is then evaluated alongside all other structures to identify the optimal description of the event.

A simple transitive event involving two participants illustrates how the model operates. The example sentence below (4.51), taken from the film script materials, describes a strongly transitive event, caused by an animate participant and affecting an inanimate one – a murderous alien predator has cut the power in a base in which he is battling a group of soldiers led by Ripley.

(4.51) **Ripley**: They cut the power.
Hudson: ... How could they cut the power, man? They're animals!

For this type of situation, English only lexicalises verbs that have the causer in the subject position (e.g. *turn off*, *interrupt*, *cut*), but for the purposes of illustration a verb that takes the causer in other positions will also be considered. The participants have the following feature values:

(4.52) *they*: causal 1; volitional 1; aware 1; physicallyAffected 0; psychologicallyAffected 0; negativeOutcome 0; possessor 0; given 1; saliency 0.5; shortness 1

(4.53) *the power*: volitional 0; volitional 0; aware 0; physicallyAffected 1; psychologicallyAffected 0; negativeOutcome 1; possessor 0; given 0; saliency 0.21; shortness 0.5;

In this example *They* represents a perfect example of ProtoAgent, being strongly causal, volitional and aware, and so is given the suitability score of 1 to this role. It has none of the qualities of ProtoPatient (score 0), but is a marginal ProtoExperiencer, being volitional and a beneficiary, but not affected (score 0.46). Similarly *the power* is a perfect ProtoPatient with similarity score 1, being physically affected in a negative way, and has none of the properties of ProtoExperiencer (0) or ProtoAgent (0).

In terms of proto-positions, *they* is more prominent, being sentient (assigned the value 0.5 as a non-human living being), given and short. Accordingly its suitability score for the first position is high at 0.66, and for the post-verb position is lower at 0.43. *The power* is considerably less prominent – it is new, longer, and as a static concrete object, comes further down on the saliency hierarchy (its score for the first argument position is 0.39, and for the second is 0.18).[29] Since there are only two arguments, the third proto-position is not entertained.

The possible active argument structures are as follows (4.54). Aggregate suitability measures for the structures are given in brackets. The as-yet unselected verb that takes a causer subject is represented by TURN-OFF, and its (unlexicalised) affected subject counterpart by BE-TURN-OFF. Indirect object arguments take a preposition, but this model does not have a mechanism for choosing it, so these are notated as PREP.

[29] It is arguable whether *the power* is new or given in this context. The definite article suggests givenness, but the fact that it has been cut is new information. The alternative in which the power is given, is treated below.

(4.54) a. they$_{\text{ProtoAgent}}$ TURN-OFF the power$_{\text{ProtoPatient}}$ [0.76]
They cut the power

b. they$_{\text{ProtoAgent}}$ TURN-OFF PREP the power$_{\text{ProtoExperiencer}}$ [0.51]
They cut at(?) the power

c. the power$_{\text{ProtoAgent}}$ BE-TURN-OFF PREP they$_{\text{ProtoExperiencer}}$ [0.27]
The power ??? ??? them

d. the power$_{\text{ProtoAgent}}$ BE-TURN-OFF they$_{\text{ProtoPatient}}$ [0.15]
The power ??? them

Among these candidates, the optimum expression according to the model is the structure (4.54a), which corresponds to the attested form *they cut the power*. Following formula (4.42) *they* has a suitability to the grammatical subject (canonical subject, in preverb position) of 0.83 (average of ProtoAgent fit of 1.0 and ProtoPreverb fit of 0.66); and *the power* a suitability of 0.70 to the grammatical direct object (average of ProtoPatient fit of 1.0 and ProtoFirstPostverb fit of 0.39). The aggregate suitability score of 0.76 for the whole structure, based around a verb such as *cut*, *interrupt* or *turn-off* is the average of these two figures (4.41).

The second most preferred structure (4.54b) characterises *the power* as an indirect object of some sort, and the choice of a preposition such as *at* would imply a reduced level of affectedness of the power. Variants in which the verb that lexicalises the meaning "to be cut" (represented with question marks) are much less preferred (4.54c,d), as the sentient causal participant *they* is characterised as a ProtoPatient, and the inanimate affected participant *power* is assigned the ProtoAgent role.

Non-canonical orderings can also be considered, such as the passive, taking canonical direct object arguments and promoting them to the grammatical subject position.[30] This is implemented in the model by giving the ProtoAgent argument the ProtoFirstPostverb position, and the ProtoPatient the ProtoPreverb position. The passive counterpart of (4.54a) is also reasonably well-formed (4.55a). Unlikely configurations such as (4.55b) can also be entertained – the passive variant of (4.54d) which uses a verb meaning "to be cut".

(4.55) a. the power$_{\text{ProtoPatient}}$ WAS TURNED-OFF by they$_{\text{ProtoAgent}}$ [0.65]
The power was cut by them

b. they$_{\text{ProtoPatient}}$ WERE BE-TURNED-OFF by the power$_{\text{ProtoAgent}}$ [0.26]
They were ??? by the power

The model is also sensitive to context effects. If *power* is given in the discourse (e.g. in response to the question "what happened to the power supply?"), the active form (4.56a) becomes slightly less appropriate (since the postverb position prefers new entities), and the passive form (4.56a) becomes slightly more appropriate. Both of these configurations have approximately equal suitability measures of 0.71, as does the truncated passive form in (4.56c). In this last case, only one participant

[30] Other marked structures of topicalisation and *it*-clefts are addressed below in the evaluation section.

is realised as an argument, so the suitability of the clause as a whole is equal to the suitability of its sole argument.

(4.56) a. They cut it [0.71]

b. It was/got cut by them [0.71]

c. It was/got cut [0.71]

Finally, the event can be considered as a candidate for conjoining arguments. Since English has a structural requirement for a subject, and there are only two potential arguments, they would both have to appear as ProtoAgents. While participant *they* is most suitably a ProtoAgent (1.0), *power* is a ProtoPatient (1.0), and their mutual suitability (or similarity) relative to ProtoAgent is low (0.0). Since the nature of each participant's involvement is so different, the model predicts that this event is not suitable for argument conjoining.

4.3.4 Evaluation

This evaluation takes two forms. Firstly, a series of situation types that allow multiple argument structure realisations and that have received attention in the linking literature are examined to see if the measure of structure suitability yielded by the model broadly predicts clause productivity. The phenomena examined are the predicates such as *meet* or *kiss* that have balanced participants that may be conjoined; the *load/spray* diathesis; the dative and benefactive diatheses; situations of giving (*give/get*); situations of spoken communication which are variably lexicalised as *tell* or *speak*; commercial transaction situations (*buy/sell/pay*); and less common orderings such as passivisation, topicalisation or *it*-clefts. Some minor variations of features of involvement and discourse will be also manipulated.

Secondly, the measures yielded by the model are compared to the acceptability measures submitted by participants in the dative/benefactive and passive diathesis experiments (sections 3.2.2 and 3.2.3). Since these sentences come in minimal pairs, the effect of differences in argument structure are isolated from those due to extraneous factors (e.g. lexical complexity; plausibility; elegance of expression; register). If a strong correlation can be found between the relative acceptability of diathesis variants, and their relative suitability according to the model, this would suggest that it is modelling perceived wellformedness.

The first diathesis to be considered is that of a balanced *meeting* situation. If both participants have arranged to meet, and neither can be viewed as being more of an instigator than the other, both will have a suitability of 1 to the ProtoAgent (being volitional, aware, causers), and a mutual suitability of 1 (relative to ProtoAgent – see 4.50). Since this mutual suitability is equal to their respective suitabilities, conjoining is licenced. Such a sentence (4.57a) receives a higher score than the Subject–Verb–IndirectObject configuration (4.57b), since that entails putting an Agent-like participant (*Jane*) in a position associated with Experiencers. The Subject–Verb–Object configuration (4.57c) comes third, since the participant *Jane* is realised in a position (the direct object) associated with Patients.

(4.57) a. Sarah and $\text{Jane}_{\text{subj}}$ met [0.88]

b. $\text{Sarah}_{\text{subj}}$ met (with) $\text{Jane}_{\text{iobj}}$ [0.71]

c. $\text{Sarah}_{\text{subj}}$ met $\text{Jane}_{\text{dobj}}$ [0.62]

If, on the other hand, two friends meet by accident, the coordinated structure is no longer preferred:

(4.58) a. $\text{Sarah}_{\text{subj}}$ bumped into $\text{Jane}_{\text{iobj}}$ [0.48]

b. Sarah and $\text{Jane}_{\text{subj}}$ bumped into each other [0.45]

And if one of the participants has more salience, in the extreme case by being the speaker, it would be expected to have privileged access to the subject. However, the model in its present form has a similar preference for salience in both the preverb and first postverb positions. Though it correctly predicts that the conjoined variant (4.58c) is no longer preferred, it does not differentiate between (4.58a) and (4.58b).

(4.59) a. I_{subj} bumped into $\text{Jane}_{\text{iobj}}$ [0.53]

b. $\text{Jane}_{\text{subj}}$ bumped into me_{iobj} [0.53]

c. Jane and I_{subj} bumped into each other [0.50]

The *load/spray* diathesis has received continuing attention in the literature (e.g. Partee, 1965; Beavers, 2004), regarding questions of the degree of affectedness attributed to each argument position. In this model, a neutral reading of the situation of hay being loaded onto a truck, in which neither the truck nor the hay is more affected than the other, nor more salient than the other, results in equal scores for both variants:

(4.60) a. I_{subj} loaded the hay_{dobj} onto the $\text{truck}_{\text{oblq}}$ [0.61]

b. I_{subj} loaded the $\text{truck}_{\text{dobj}}$ with the hay_{oblq} [0.61]

In this analysis the prepositional objects are not indirect objects, but oblique phrases, so they do not hold a proto-role, but do hold a proto-position. Analysing them as an indirect object lowers the aggregate suitability scores of both sentences to 0.53, as neither the *truck* nor the *hay* have many attributes of ProtoExperiencer.

Changing the level of affectedness of one of the participants, makes the choice between diatheses clear (a value of 0.5 is used to denote 'partial' affectedness, and 1.0 to denote 'holistic' affectedness). In the case that the truck is holistically affected, and the hay only partially affected, the model predicts (4.61a) as the preferred realisation, rather than (4.61b).

(4.61) a. I_{subj} loaded the $\text{truck}_{\text{dobj}}$ with some of the hay_{oblq} [0.61]

b. I_{subj} loaded some of the hay_{dobj} onto the $\text{truck}_{\text{oblq}}$ [0.56]

But discourse effects can have a similar results. In a case in which both *truck* and *hay* are partially affected, but only the *hay* is new to the discourse, the variant in which the *truck* is realised as a direct object is again preferred.

(4.62) a. I_{subj} loaded the $\text{truck}_{\text{dobj}}$ with some hay_{oblq} [0.66]

b. I_{subj} loaded some hay_{dobj} onto the $\text{truck}_{\text{oblq}}$ [0.60]

The benefactive diathesis is also modelled correctly. In cases in which the indirect object is a volitionally involved beneficiary, and all participants are given in the discourse, both NounPhrase-PrepositionalPhrase and NounPhrase-NounPhrase configurations are similarly suitable (an analysis of 4.63b in which *for you* is analysed as an oblique yields a lower score of 0.65).

(4.63) a. I_{subj}'ll fetch you_{iobj} your $slippers_{dobj}$ [0.73]

b. I_{subj}'ll fetch your $slippers_{dobj}$ (for) you_{iobj} [0.71]

If however one of the participants is new to the context the variant that places this argument at the tail of the clause is preferred (4.64a). However, manipulating argument length, as in (4.65), has little appreciable effect, as the strength of association found with this feature was very weak (see table 4.2).

(4.64) a. I_{subj}'ll fetch you_{iobj} some $slippers_{dobj}$ [0.82]

b. I_{subj}'ll fetch some $slippers_{dobj}$ (for) you_{iobj} [0.74]

(4.65) a. I_{subj}'ll fetch you_{iobj} your new cosy slippers that you got on $holiday_{dobj}$ [0.73]

b. I_{subj}'ll fetch your new cosy slippers that you got on $holiday_{dobj}$ (for) you_{iobj} [0.70]

An unsuccessful benefactive transformation, such as (4.66c) is a less acceptable description than either NounPhrase–PrepositionalPhrase configuration. Of these two (4.66a,b), that which analyses the phrase *two thousand years* as an oblique is preferred.

(4.66) a. $They_{subj}$ been making $statues_{dobj}$ (for) two thousand $years_{oblq}$ [0.66]

b. $They_{subj}$ been making $statues_{dobj}$ (for) two thousand $years_{iobj}$ [0.60]

c. $They_{subj}$ been making two thousand $years_{iobj}$ $statues_{dobj}$ [0.59]

A similar pattern obtains for the dative diathesis. A sentence that was judged as not acceptable in the double object configuration (*my friend ... took me to that video arcade*) is most suitable in a Subject–DirectObject–Oblique (0.68) configuration, and least successful in the Subject–IndirectObject–DirectObject configuration (0.50). One that was judged acceptable (*Remind me to send Mr.Boeing a thank you note*), had similar scores for both configurations (0.75 and 0.77).

A dative situation in which the Donor is not realised can also be considered. In (4.67) the first double-object variant is more successful.

(4.67) a. Give me_{iobj} a $present_{dobj}$ [0.78]

b. Give a $present_{dobj}$ to me_{iobj} [0.66]

In English these subjectless structures are limited to imperative uses. A more general alternative is to either use a passive (4.68a) or use another verb that lexicalises the Recipient as the canonical subject (4.68b). While the passive is perhaps more suitable than might be expected (as has already been mentioned, no discounting of suitability is imposed for long or unwieldy constructions), this indicates that the ProtoExperiencer–Verb–ProtoPatient configuration (4.68b) is also a desirable description worth lexicalising.

(4.68) a. I_{subj} was given a $present_{dobj}$ [0.78]

b. I_{subj} got a $present_{dobj}$[0.65]

Situations of spoken communication can also reveal something about potential and extant lexicalisation patterns. *Talk* or *speak* often take a prepositional object rather than a direct object (though not all such verbs do – consider *ask* or *phone*). This is reflected in the ranking of the configurations given by the model (here the token COMMUNICATED is used to represent the class of verbs that take a Speaker subject, and BE-COMMUNICATED those that take an Addressee subject – *she* is the Speaker and *I* is the Addressee):

(4.69) a. She and I_{subj} COMMUNICATED [0.79]

b. She_{subj} COMMUNICATED (to/with) me_{iobj} [0.76]

c. I_{subj} was COMMUNICATED (to/with) by her_{oblq} [0.76]

d. She_{subj} COMMUNICATED me_{dobj} [0.67]

e. I_{subj} BE-COMMUNICATED (from) her_{iobj} [0.56]

f. I_{subj} BE-COMMUNICATED her_{dobj} [0.53]

The first three configurations are consistent with verbs such as *speak* and *talk*, while (4.69d) would suit verbs such as *ask* or *tell*. In the final two variants, the model considers the possibility of expressing the communication with a reversed canonical linking. The low scores these two variants receive are consistent with the fact that English-speaking communities do not seem to feel the need to lexicalise a verb meaning *to be spoken to*.

In a commercial transaction where *Tom* is the purchaser, and *Mary* the vendor, the model predicts the following ranking of argument configurations (neither vendor nor purchaser is treated as the sole instigator):

(4.70) a. $Mary_{subj}$ sold Tom_{iobj} a new car [0.74]

b. $Mary_{subj}$ sold a new car (to) Tom_{iobj} [0.67]

c. #Tom_{subj} bought $Mary_{iobj}$ a new car [0.67]

d. Tom_{subj} bought a new car (from) $Mary_{iobj}$ [0.60]

e. ?Tom_{subj} relieved? Mary (of) a new car [0.53]

f. *$Mary_{subj}$ sold Tom (of?) a new car [0.50]

g. A new car_{subj} ???? Mary (to) Tom [0.33]

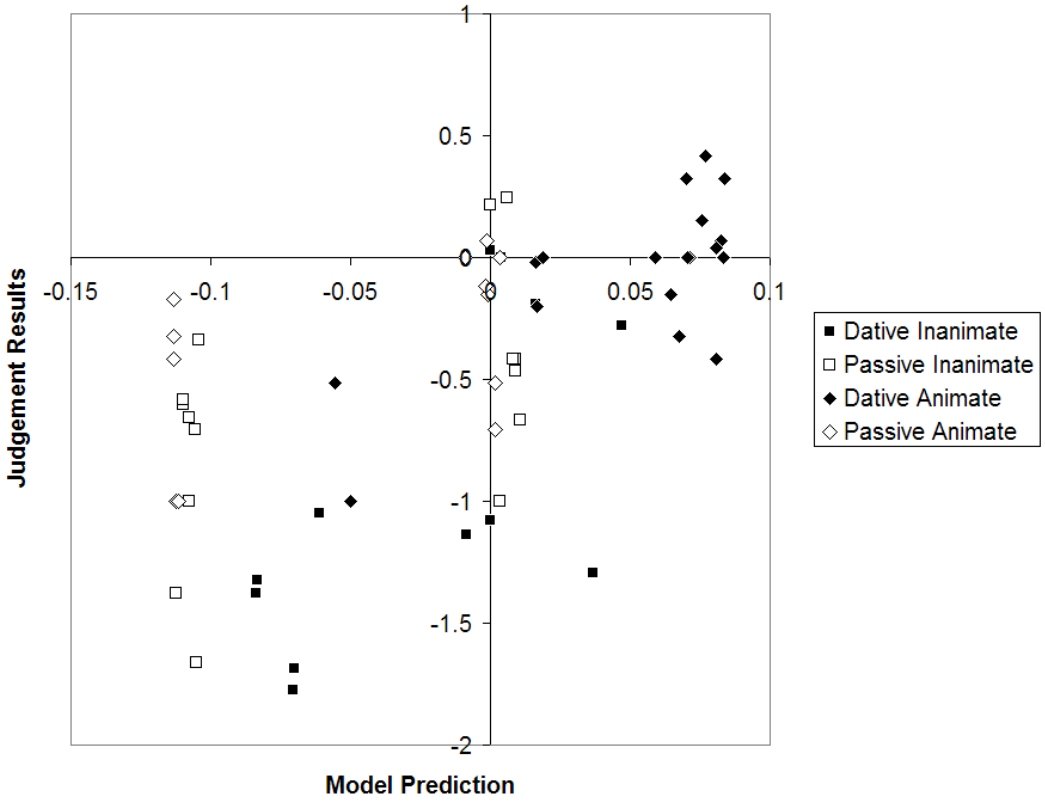
h. A new car_{subj} ???? Tom (from) Mary [0.22]

Again, the model seems to prefer forms that are lexicalised to those that are not. The four top-ranked candidates are wellformed sentences, though this interpretation of (4.70c) is blocked by the benefactive construction (*Tom bought a car for Mary*). The last two candidates receive much lower scores. Assuming that the model simulates desired argument structure realisations, verbs are unlikely to be lexicalised for structures such as (4.70g,h) that speakers never, or rarely want to use.

Verbs of opinion are another case in which similar meanings are expressed with multiple lexemes which have distinct argument structures. The sentence (4.14a) on page 156 can be construed in two ways. It can be taken as a description of a change of opinion (on the part of *India's software industry*) in which the *scheme* is causal (and so has elements of ProtoAgency), or as a description of a state in which the *scheme* has no strong involvement in the event, rather being an unaffected object of cognition. In the first dynamic causal reading, the model strongly prefers the attested construction based around the verb *please* (*the scheme ... does not please ... industry*, 0.41), or a passive construction (*industry is not pleased by the scheme*, 0.39) over an active sentence using *like* (*industry does not like the scheme*, 0.24). In the second stative reading however, all three variants are approximately of the same suitability (with scores of 0.27, 0.25 and 0.24 respectively).

Finally, an adequate theory of English argument structure should explain less common orderings such as the passive, topicalisation and *it*-clefts. As was shown in the last section, differences in argument structure can make a passive structure (4.56b) as suitable as its active counterpart (4.56a)

Figure 4.2: Judged diathesis success scores and model predictions



(in this case, *it* was given in the context and provided coherence with a preceding context). The cleft and topicalisation in (4.71) can be considered to place *the power* and *they* in the roles of direct object and subject (as in the canonical active sentence), and both arguments can be seen to occupy a preverb position (relative to *cut*). This yields a high score of 0.77 (the equivalent active and passive forms each scored 0.71), despite the clumsiness of (4.71a) (relative to the active and passive forms, which are both available), and the illformedness of (4.71b).

(4.71) a. It was the power$_{\text{dobj}}$ that they$_{\text{subj}}$ cut [0.77]

b. ?The power$_{\text{dobj}}$, they$_{\text{subj}}$ cut [0.77]

Overall, these comparisons demonstrate that in the majority of cases the model correctly predicts the preferred descriptions of an event. Relative to the set of candidate utterances for a given situation, the suitability measures give an approximate ordering of appropriateness. However, it does not seem that these measures are comparable across situations. For example, it is not clear that the top ranked realisation for a planned meeting *Sarah and Jane met* (4.57a) which was given a score of 0.88, is considerably more acceptable than the optimum realisation of an unplanned meeting *Sarah bumped into Jane* (4.58a) which only received a score of 0.45. Rather the absolute suitability scores reflect similarity to a prototypical transitive event (the *liking* situation just discussed is another example of an event that is far from this prototype, and it received correspondingly low scores, despite being wellformed). But the intervals between suitability scores may provide a prediction for relative judged acceptability.

For the evaluation against the acceptability data, suitability scores were calculated for vari-

ants of test and filler sentences from the English language acceptability experiment. For the dative/benefactive sentences the suitability score for the marked double object construction was subtracted from the suitability score of the unmarked NounPhrase-PrepositionalPhrase construction. For the passive/active sentences the score for the passive variant was subtracted from that of the canonical active. The resulting scores are an indication of the relative suitability of the non-canonical variants. These were then compared to the difference scores of the dative/benefactive and active/passive materials from the acceptability experiments, which also represent the relative success of the diatheses.[31]

As figure 4.2 shows, there is a clear positive correlation between acceptability judgements and model predictions when diathesis variants are compared. The linear correlation coefficient r over all sentences is 0.63, which means that the model is explaining approximately 40% human acceptability judgements of these diatheses (r^2, the coefficient of determination). The trend among the passive sentences alone is similar at $r = 0.62$. Among the dative/benefactive sentences the trend is stronger ($r = 0.80$, $r^2 = 0.64$), and the detected effect of animacy on diathesis success is clearly apparent, with the black diamonds representing sentences with animate indirect objects clustering in the positive quadrant and the black squares representing sentences with inanimate indirect objects principally in the negative quadrant. Thus, this comparison suggests that while the output measure of appropriateness does not predict absolute acceptability, it can successfully model the relative acceptability of candidate verbalisations of a single event.

4.4 Implications

The model just described is to some extent arbitrary, and fails to predict relative suitability of structures in some cases. In particular it does not correctly predict orderings that invert the canonical ordering (passive, clefts, topicalisation). This could be addressed in part by adding a term to the computations that rewards brevity of expression. But there also appear to be other properties, not necessarily of participants themselves, that contribute to the licensing of such structures, such as telicity for the English passive, negative affect for Chinese *bei* and a restriction to concrete events for the Chinese *ba* construction. Further adjustments that would have to be made to the model to provide a comprehensive description of argument structure realisation would be to reward the use of common vocabulary, and to introduce a model of focus relations (which would in turn inform a model of intonation realisation). An extension of the model to explain the relative orderings of obliques, and their choice of preposition or other connective could also be considered.[32]

However, despite its numerical and computational simplicity[33], and the very small number of sentences used to derive its parameters[34], the model can account for a wide range of phenomena, both

[31] There were 30 dative/benefactive sentences, of which 24 were test sentences (each a NounPhrase-PrepositionalPhrase original compared to its constructed double object variant; half had animate indirect objects) and 6 filler sentences (double object originals compared to their NounPhrase-PrepositionalPhrase variants; all had animate indirect objects). There were 38 passive sentences of which 30 were test sentences (active original, passive altered) and 8 were fillers (passive original, active altered). Of these 38, half had an animate canonical direct object. The acceptability difference scores for the filler sentences represented the success of the diathesis *out of* the non-canonical form, so their sign was switched to make them directly comparable to the other scores.

[32] Here, the localist notions of Source and Goal might be usefully employed in the domains of Time, Space, and Contingencies (differentiating between motivations that precede and follow an event).

[33] While the model might seem complex compared to models that use a simple role hierarchy, it is not computationally intensive when compared with the general cognitive load of lexical selection, where choices must be made between thousands, rather than tens of possible configurations.

[34] Some of the false predictions of the model might be addressed by using a larger base of materials for parameterisation. For example the failure to predict that *I met Jane* is generally preferable to *Jane met me* is due to the fact that association between saliency and the preverb and postverb positions were roughly equal. The literature suggests that the association with the preverb position should be stronger. The observed effect of argument heaviness was also different from that expected.

in terms of the choice of expressions during production, and their relative perceived acceptability. The model (or any similar model that shared its architectural assumptions of no primacy of any features; competition between all potential realisations; and properties based on event construals) can explain argument structure and verb lexicalisation patterns without reference to subtle differences in the lexical semantics of verbs. For example the causal connotation of *please* can be viewed as a consequence of general principles of the grammar (that subjects tend to be Agentive), rather than a distinctive feature of the verb relative to its counterpart *like*. The integration of elements of discourse into the model allows it to explain selection between truth-conditionally equivalent diathesis pairs, and to account for the existence of verbs with atypical argument structures (since we often want to talk about Recipients, and they are by definition animate and therefore salient, the lexicalisation of *receive* provides an economical and more elegant alternative to *X was given Y*).

The features appealed to do not need to claim any primitive status, and are generally of a common-sense nature (i.e. one can imagine asking naive informants *was X physically affected by event Y*; whereas notions such as Dowty's incremental theme would need extensive introduction). Further, since all structures can be generated from first principles, there is no requirement to store argument structure frames (as in Goldberg, 1995), other than for atypical verbs such as *undergo*. And this also provides for novel argument structure usages of existing verbs (e.g. Goldberg's *Sally sneezed the napkin off the table*) and novel verbs (e.g. to be *gold-watched by the employees* – see page 3). This ability to generate all structures may be even more compelling for languages that do not have as rigid an argument structure as English.

As mentioned at the outset, this work did not assume any particular theory of syntax. The type of model proposed here is incompatible with many such theories for two reasons. It is based on prototype concepts and continuous categories rather than on atomic symbols. And it requires interaction between elements of syntax, semantics and discourse, which are seen as modularly separate by some theories (most prominently in the generative tradition). However a monostratal theory such as HPSG (Pollard and Sag, 1994) would be well suited to representing the resulting structures (unlike RRG or LFG for example). Though that framework does not make grammatical functions explicit, the combination of obliqueness ordering in the SUBCAT feature, the inclusion of prepositions and subordinators, and the indexing of semantic relations (such as Giver, Loved, etc.) represent both serial position and canonical grammatical function in an alternative form. A more extensive reworking might replace those semantic relations with the proto-roles of this model. It is less obvious how the competition model could be operationalised in the symbolic framework of HPSG, other than using a ranking procedure along the lines of Optimality Theory.[35]

One can also speculate on the extent to which a model of this type might generalise to other languages. Certainly differences in parameter values would be expected. But it may be that the tendencies seen in the English model are general principles of cognition and communication – that is that speakers view causes as prior to effects (cf. Langacker, 1990), and order them accordingly in sentences; that speakers tend to give prominence to participants with which they empathise or identify; and that speakers place given information early in the clause to aid the listener in linking a sentence into the preceding discussion. On such a view, languages would be expected to be relatively uniform in their expression of strongly transitive events, placing Agents before Patients (since they involve a clear asymmetry of cause and effect, and because causers are typically animate, and so have more anthropocentric saliency). In situations in which considerations of involvement and discourse

[35] Approaches that resolve conflicting constraints in feature structures and other logic-based formalisms, or introduce graded satisfaction of such constraints, include fuzzy logics and paraconsistency (see e.g. Vogel and Cooper, 1995).

are not in concert (e.g. detransitives, or psych verbs), one might expect arbitrary variation in the mechanisms that languages use to resolve such conflicts.

Indeed, a difference in the grammaticalisation of these putative cognitive universals can be seen. English is a very strongly nominative language, in that its subject is obligatory, must be placed in the prominent preverb position, takes verb agreement, and is associated with causation. Thus it is a good exemplar of Keenan's universal subject.[36] Word order is rather fixed in English, so non-transitive events are often expressed by pairs of verbs that allow the speaker to take one or other perspective (e.g. *fear/frighten*), or by limited verb diatheses (e.g. the detransitive, or passive).

Spanish is also a nominative language, but has a freer word order, and also a mechanism of duplicating arguments with a pronoun (a clitic). This allows the speaker to "hedge their bets" in a manner of speaking. As discussed, in the detransitive construction the sole ProtoPatient argument is given the grammatical role of subject (as Spanish requires one), but is often placed in the postverb position (associated with affectedness), while a clitic pronoun remains in the preverb position. Similarly in *liking* constructions, an indirect object pronoun is used to give the Experiencer the discourse prominence of the preverb position, without any connotations of causation.

An ergative language such as Tibetan might be expected to have a different ordering, since its absolutive case (which is identified with affectedness) has the syntactic prominence that is given to subjects in nominative languages (e.g. in Tibetan, the sole argument in an intransitive takes the absolutive case, regardless of the nature of its semantics). However, as the following sentences suggest (Goldstein et al., 1991, p.57), canonical ordering does still place agent-like participants earlier in the clause.

(4.72) dɔ̲ɔ̲jeè$_{ergt}$ kāŋba$_{absl}$ ña̲mso che̲shinyöö̀$_{verb}$
dorje house repair is-doing
'Dorje is repairing the house'

(4.73) pū̲ndzo-ki$_{ergt}$ go̲mbar$_{absl}$ chō̲njɛɛ shu̲shinyöö̀$_{verb}$
phuntso to-monastery religions-visit is-doing
'Phuntso is making a religious visit to the monastery

In Basque also, new information tends to be placed towards the end of the sentences (Ortiz de Urbina, 2002, p.3), and the neutral clause order puts the agent-like ergative case before the patient-like absolutive (Arregi, 2001, p.1; Laka, 2008, sec. 1.0).

(4.74) Martinek$_{ergt}$ egunkariak$_{absl}$ erosten$_{verb}$ dizkit$_{aux,dative}$
Martin newspapers buying is-to-me
'Martin buys the newspapers for me'[37]

In fact, very few ergative languages place the patient before the agent in the unmarked form (Baker, 1997, p.7), following a general pattern of the great majority of human languages (of 1059 languages with a dominant clausal order surveyed by Dryer (2005), only 39 place patient-like arguments before agent-like arguments).

[36] Keenan's (1976) typological characterisation of the category of subject lists the typical, but not necessary, attributes under four headings: Autonomous Reference, Case Marking, Semantic Role and Immediate Dominance. Autonomous Existence covers questions of independent existence (whether the referent of an argument is created or destroyed in process of an event), indispensability (whether an argument is optional) and autonomous reference (i.e. definiteness), while the Semantic Role of subjects is usually of an agent or causer type. If a language has case marking then its subject tends to be one of the grammatical roles that uses it, and the subject usually comes early in the sentence.

[37] Wikipedia entry on the Basque language [http://en.wikipedia.org/wiki/Basque_language], viewed August 2006.

Chinese might be considered a "marginally" nominative language. Its subject is associated with causation, and appears by default in the prominent topic position at the head of the clause. However there is no case marking and no agreement with the verb. Further, the subject can be freely dropped, and any argument or adjunct can move ahead of it into the topic position. This has lead some to question if Chinese can be said to have a subject at all (see footnote 28 on page 80). However the subject's default position is before the object (following the cause before effect schema); and given information does tend to precede new.

Dryer (1997) describes the Philippine language Cebuano to justify his claim that their is in fact no *universal* category of subject. In this language, the grammatical marking of participation seems to be completely decoupled from that of discourse considerations. Topics and non-topics are signalled with the markers *ang* and *sa* respectively (4.75). The semantic role of the topic argument is indicated separately by inflecting the verb with the morphemes *mi* or *gi* (*ibid.* pp.131-132) (the role of the non-topic argument can then be inferred).

(4.75) a. mi-palit ang babayi sa saging
ACTOR.FOCUS-buy TOPIC woman NONTOPIC banana
'The woman bought the bananas'

b. gi-palit sa babayi ang saging
GOAL.FOCUS-buy NONTOPIC woman TOPIC banana
'The woman bought the bananas'

Dryer (1997, pp.122,139) rather makes a claim that is neither universalist nor Whorfian: that the similarities seen between grammatical roles across languages is a consequence of common "cognitive and functional principles", and differences are due to how "competition among motivations is resolved". Comrie and van den Berg, 2006, report seven different distributions of non-transitive Experiencer constructions, just among Daghestani languages. Croft (1991, p.151) also observes that languages that have more rigid constraints on the selection of grammatical relations (unlike English, which has many diatheses – see Levin, 1993) tend to allow freer movement of arguments.

So in this way it may be that the relative infrequency of the passive constructions in Spanish and German (relative to English – Zagona, 2002, p.40; Durrell, 2002, pp.476-480) is a consequence of the existence of the more economic alternatives of argument movement. Similarly, the general processes of topicalisation and dropping in Chinese allow that language to produce structures that are functionally equivalent to the English passives and detransitive, without recourse to restructuring of grammatical relations, or particular morphological marking.[38] However this does leave unanswered the question of why apparently arbitrary restrictions apply to some argument structure configurations (such as concreteness for the Chinese *ba* and adversity for the Chinese *bei*; or English verbs like *destroy* that resist the detransitive form). Particularly in English, it would be very desirable to have a universally productive marking mechanism to signal that serial order has been altered while canonical grammatical function assignment remains unchanged, but all the available structures (passives; topicalisations; clefts) are unwieldy or limited in their applicability.[39]

If it is indeed true of languages generally that information structure, saliency and involvement pattern in the way seen here (given salient actors tend to appear early in the clause; new less-

[38] It is also interesting to note that English, whose general mechanisms for object fronting are the rather unwieldy passive and topicalisation, has common verbs in both configurations: Stimulus-subject (*please, satisfy, make happy*) and Experiencer-subject (*fancy, enjoy, appreciate, like*). In Spanish on the other hand Stimulus-subject verbs predominate (e.g. *caer bien, gustar, apetecer*) over Experiencer-subject verbs (*agustar, dar gusto*), as the dative clitic pronouns such as *me* allow Experiencer-object prominence, without changing the assignment of grammatical functions.

[39] Dowty (2001) speculates that such restrictions, or "filtering effects" may be a general characteristic of diatheses.

consequential patients later), it may be that the cross-linguistic variation seen in argument structure is a relatively superficial phenomenon – largely a consequence of arbitrary syntactic strategies adopted to deal with less prototypical situations. Languages would agree on the representation of "core" situations and differ on the treatment of "marginal" ones.

This would amount to constrained and continuous typological variation of argument structuring, on the basis of some universal principles. Such variation is known to operate in the realisation of vowels across languages (Liljencrants and Lindblom, 1972). Also, in the lexicalisation of colour terms, Regier et al. (2007) demonstrate that though inventories vary from language to language, they all constitute a near optimal partition of the colour space (following perceptual principles of maximum similarity within categories and minimum similarity between categories). Though very different in its details, such a theory is also similar in spirit to Chomsky's Principles and Parameters (1981).

To answer these questions, more work would be required. The English model could be improved by introducing factors of economy and of rhetorical focus. A larger corpus of sentences would provide a better basis for parameterisation (covering a wider range of registers and domains), and the coding of features could be verified more extensively, or pre-annotated text could be used (e.g. the FrameNet or WordNet corpora). Materials that included examples of passives, topicalisations and clefts might allow the model to improve its predictions of these argument structure inversions. Finally, such a procedure should be replicated in several other languages, including ones that are ergative, and ones that have a dissociation between the coding of participation and information structure. German and Chinese would be good initial candidates, since acceptability data has been gathered, and the effects of semantic and pragmatic factors in their argument structures have been confirmed. Though both are nominative languages, they are also structurally interesting. German differs from English in having comprehensive case marking, and relatively free movement of arguments, and Chinese exhibits a partial decoupling between the encoding of participation and discourse properties.

4.5 Final Remarks

The work described in this thesis did not take any detailed model of processing or form of representation as a starting point, which would then be falsified or refined by means of experiment. Rather the aim was to trawl for linguistic factors that could be shown to have a significant determining effect on the realisation of argument structure in several languages.

The experimental results suggested that a comprehensive description of English linking patterns is not possible without taking pragmatic and discourse factors into account (in addition to the better studied effects of syntax and semantics). Furthermore, they suggested that the same may hold for German and Chinese, which is also pretheoretically plausible, since both employ argument movement more widely than English. And no clear categorial difference was apparent between the magnitude of acceptability effects seen for syntactic and other violations (as has been suggested by Chomsky, 1964 and Sorace and Keller, 2005 among others).

In addition, a wide range of argument structure phenomena that have been widely discussed in the literature (including the dative/benefactive diatheses, the *load/spray* diathesis, and causal/noncausal psych-verb pairs such as *like/please*) could be explained by appealing only to language general linking principles, and without reference to the lexical semantics of individual verbs. A consequence of this is that competition can operate between candidate utterances built around different verbs, and

that the speaker's construal (that is their conceptual description of the event, in whatever form that takes) may select the form of verbalisation, rather than vice-versa.

This leads to what is perhaps a more functionally compelling account of sentence generation than rigid classical models (e.g. Chomsky, 1981; Levin and Rappaport Hovav, 2005), that assume lexicalisation of the predicate precedes or dominates other choices of verbalisation. Rather speakers already have a construal of an event in mind, based on their experience or understanding of it. On this basis they select the most appropriate verb lexicalisation and clause structure offered by their language, subject to language-general violable constraints on the interpretation of grammatical function and sentence position.

Appendix A

Notations Used

Marks of wellformedness

***** Syntactically illformed

Semantically illformed

@ Semantically divergent from counterpart

% Incorrect parse

Grammatical Notations

absl Absolutive case

attr Attributive marker

aux Auxiliary verb

dobj Direct object

ergt Ergative case

intr Interrogative marker

iobj Indirect object

neg Negative Affect

neu Neutral Affect

oblq Oblique prepositional or nominal argument

pfv Perfective aspect

pos Positive Affect

pred Predicative phrase

prep Preposition

refl Reflexive pronoun

subj Subject

Appendix B

Experimental Materials

Instructions for English Diathesis Experiment

Instructions for Film Script Questionnaire[1]

In this questionnaire you will see 30 short extracts from film scripts - it should take you about 20 minutes to complete. Some of the quotes will seem perfectly OK to you, but others won't. Your task is to decide how English-sounding a particular quote is. Imagine you are helping a friend learn English. If your friend used the same sentence, would it sound like a native talking (ignoring pronunciation)? Or would it sound strange or unnatural?

Since these are film dialogues, the style is quite casual. You *don't* have to worry about whether a quote is bad because of punctuation, slang, hesitations, or repetitions and changes the speaker makes as they go along. For example, "I gotta go to the ... uh ... like ... my parents' place." would be fine, because English speakers often say things like that. Also, don't worry about any grammar rules you may have been told in school - there are no right or wrong answers. And don't spend too long on any quote - it is your *first impression* that we are interested in.

How to score the quotes

You will have to give each quote a score. The first quote you see will be the "reference quote", and you will score all other quotes compared to this. You can choose any number you want for the reference quote. Then, for each following quote, use a bigger number if you feel it is more English-sounding, and give a smaller number if you think it is strange or unnatural.

For example, the first sentence below is okay English, but maybe a little awkward - let's give it a reference score of 100.

1. The mat was what the cat sat on

You might feel that the second sentence is twice as good, because it sounds much more natural. Then you might give it a score of 200. [that is 100 multiplied by two]

2. The cat sat on the mat

The third sentence has a clear mistake ('sitted' instead of 'sat') - if you felt it is twice as bad as sentence 1, you would give it a score of 50. [that is 100 divided by two]

3. The cat sitted on the mat

The last sentence is hard to make sense of at all, so if you find it three times worse than the reference you could give it a score of 30. [roughly, 100 divided by three]

[1] Original can be viewed at https://www.cs.tcd.ie/Brian.Murphy/bestest/instructions.html

4. cat sat on mat the

Of course you can use any reference score you want. If you had chosen 20 for sentence 1, you might have given 40 for sentence 2, 10 for sentence 3, and 7 for sentence 4.

When you are scoring a quote we will also give you a little bit of context - a short description of the film and the scene it belongs to. You should read these first and reply 'yes' or 'no' to the question of whether you are *familiar* with the plots of the film and scene in question (you *don't* have to know the scene word for word). Then you should read the dialogue and give a score for the underlined portion of text. So, for the example question below, you should give a score based on your impression of "... they'd be left at the city ..." only.

[Film] The Empire Strikes Back, George Lucas, 1980: The Star Wars rebel alliance continue to battle the evil Empire led by Darth Vader.	"I have seen this film" Yes: [radio button] No: [radio button]
[Scene] Darth Vader goes back on his deal with Lando Calrissian (Billy Dee Williams) to let Princess Leia (Carrie Fisher) and Chewbacca the Wookie go free.	"I remember this scene" Yes: [radio button] No: [radio button]
Darth Vader: ... Take the princess and the Wookie to my ship. **Lando:** You said they'd be left at the city under my supervision. **Darth Vader:** I am altering the deal. Pray I don't alter it any further.	Score for underlined text: [textbox containing the number 150]

So, to step through this example:

- You saw the film "The Empire Strikes Back" a long time ago, and you can roughly remember what it is about, so you answer "Yes" to the question "I have seen this film".
- You can't remember the scene where Leia and Chewbacca are taken prisoner, so you answer "No" to the question "I remember this scene".
- Say the reference quote was "I get them it for half", and it was given a score of 100. You think that "they'd be left at the city" is one and a half times as good as the reference quote, so you give it a score of 150. [100 multiplied by one and a half]

When scoring you can use any range of positive numbers that you like, including decimals and fractions. You can't use zero or negative numbers. Try to use a wide range of numbers to distinguish as many levels as possible.

Remember:

- Give the reference quote any number you want (it will be repeated at the start of each page as a reminder).
- Score all the other quotes relative to this one, giving big numbers for English-sounding ones and small numbers for strange or unnatural ones.
- Try to make up your mind as quickly as possible without rushing. There are no 'right' or 'wrong' answers - it is your first impression we are interested in.
- Use as wide a range of numbers as you want.
- Remember to mark 'yes' or 'no' to tell us if you are familiar with the scene and film.
- Base your score only on the underlined portion of the dialogue.
- Please go straight through the questionnaire, without using the 'back' or 'reload' buttons on your browser.

Click below to start the questionnaire.

[Start Button]

Instructions for German Diathesis Experiment

Hinweise zum Ausfüllen des Fragebogens für die Film-Studie[2]

Im nachfolgenden Fragebogen werden Ihnen 30 kurze Textausschnitte aus Drehbüchern präsentiert, deren Bearbeitung ungefähr 20 Minuten Ihrer Zeit in Anspruch nehmen wird. Einige dieser Zitate werden Ihnen völlig richtig erscheinen, andere wiederum nicht. Ihre Aufgabe besteht darin zu entscheiden, in welchem Maß die jeweiligen Ausschnitte normales, gutes Deutsch sind. Als Entscheidungshilfe: Wenn Sie sich vorstellen, dass Sie einem Freund helfen Deutsch zu lernen und dieser den entsprechenden Satz sagen würde, würde dies (abgesehen von der Aussprache) wie von einem Muttersprachler klingen? Oder würde es komisch oder unnatürlich klingen? Da die Zitate aus Filmen stammen, ist der Stil eher umgangssprachlich. Sie brauchen *nicht* in Betracht ziehen, ob ein Satz wegen Zeichensetzung, Slang, Pausen, Wiederholungen und Satzänderungen durch den Sprecher mangelhaft ist. Z. B., “Kannste mir mal das ... äh ... ich meine ... deinen Kuli borgen?" wäre in Ordnung, weil Muttersprachler so etwas häufig sagen. Sie brauchen sich auch nicht die Grammatikregeln, die Sie vielleicht einmal in der Schule gelernt haben, wieder in Erinnerung rufen - es gibt keine richtigen und falschen Antworten. Sie sollten auch nicht allzu viel Zeit auf jedes einzelne Beispiel verwenden - was zählt und uns interessiert, ist der erste Eindruck, also Ihre spontane, unreflektierte Bewertung.

Vorgehensweise zur Bewertung der Zitate

Zum Ausfüllen des Film-Fragebogens geben Sie jedem der 30 Zitate eine Bewertung. Sie beginnen damit, dass Sie als allererstes einen Bezugssatz (Zitat 1) bewerten, der für alle nachfolgenden Zitate die Bewertungsgrundlage bildet. Dieser Bezugssatz wird gemeinsam mit der von Ihnen vergebenen Bewertung am Anfang jeder Seite zur Erinnerung eingeblendet. Sie können jede beliebige Zahl als Bewertung für den Bezugssatz verwenden. Für die weiteren Zitate verwenden Sie dann eine größere Zahl, wenn Sie der Meinung sind, dass das Zitat besseres Deutsch als der Bezugssatz ist, oder kleinere Zahlen, wenn das Zitat Ihrer Meinung nach eigenartig oder unnatürlich klingt.

Zum Beispiel ist der erste Satz unten normales Deutsch, außer dass er vielleicht etwas umständlich ist. Sie entscheiden sich, ihm 100 Punkte zu geben.

1. Der Läufer war es, worauf die Katze saß.

Der nächste Satz ist Ihrer Meinung nach zweimal so gut wie der erste, da er viel normaler klingt. Dann würden Sie ihm 200 Punkte zu geben (100 mal 2 macht 200).

2. Die Katze saß auf dem Läufer.

Der dritte Beispielsatz enthält eindeutig einen Fehler ("sitzte" statt “saß"). Wenn Sie der Meinung sind, dass er deshalb doppelt so schlecht wie der Bezugssatz ist, würden Sie ihn mit 50 Punkten bewerten (50 = 100 durch 2).

3. Die Katze sitzte auf dem Läufer.

Den letzten Beispielsatz kann man kaum verstehen, deshalb finden Sie ihn, angenommen, dreimal so schlecht wie den Bezugssatz. In diesem Fall müssten Sie ihm ungefähr 30 Punkte geben (100 durch 3 ergibt 33).

4. Die Katze saß auf Läufer dem.

Es wäre auch möglich gewesen, den Bezugssatz z. B. mit 20 zu bewerten. Dann hätten Sie Satz 2 mit 40, Satz 3 mit 10 und Satz 4 mit rund 7 bewertet.

Zusätzlich zu den zu bewertenden Zitaten geben wir Ihnen auch ein bisschen Kontext: eine kurze Beschreibung des Films und aus welcher Szene des Films das Zitat stammt. Diese lesen Sie bitte zuerst und beantworten danach die Fragen, ob Sie den Film mit seiner Handlung kennen bzw. sich an die fragliche Filmszene erinnern, entsprechend mit ‘Ja’ oder ‘Nein’. (Sie müssen die Szene *nicht* Wort für Wort auswendig können.) Dann lesen

[2]Original can be viewed at https://www.cs.tcd.ie/Brian.Murphy/bestestDE/instructions.html

Sie bitte das gesamte Zitat, bewerten aber nur den unterstrichenen Teil. Z. B. vergeben Sie im Beispiel unten die Punkte ausschließlich auf Grund Ihrer Einschätzung von “... sie unter meiner Aufsicht in der Stadt bleiben würden ...".

[Film] Krieg der Sterne - Das Imperium schlägt zurück, Irvin Kershner, 1980: Die Allianz der Rebellen führt ihren Kampf gegen das von Darth Vader geführte Imperium fort.	“Ich habe diesen Film gesehen." Ja: [radio button] Nein: [radio button]
[Filmszene] Darth Vader hält seinen Vertrag mit Lando Calrissian (Billy Dee Williams) über die Freilassung von Prinzessin Leia (Carrie Fisher) und Chewbacca den Wookie nicht ein.	“Ich kann mich an diese Szene erinnern." Ja: [radio button] Nein: [radio button]
Darth Vader: ... Bring die Prinzessin und den Wookie auf mein Schiff. **Lando:** Sie hatten gesagt, dass sie unter meiner Aufsicht in der Stadt bleiben würden. **Darth Vader:** Ich habe den Vertrag geändert. Hoffe und bete, dass ich ihn nicht noch weiter abändere.	Bewertung für unterstrichenen Ausschnitt: [textbox containing the number 150]

Eine schrittweise Beschreibung für die Bewertung dieses Beispiels:

- Sie haben den Film “Das Imperium schlägt zurück" vor langer Zeit einmal gesehen. Sie können sich ungefähr daran erinnern, worum es ging, und beantworten deshalb die Aussage “Ich habe diesen Film gesehen." mit “Ja".
- Sie können sich nicht an die Szene erinnern, in der Leia und Chewbacca gefangen genommen werden, und antworten deshalb auf die Aussage “Ich kann mich an diese Szene erinnern." mit “Nein".
- Angenommen, Sie hätten dem Bezugssatz “Ich besorg denen die für die Hälfte." 100 Punkte gegeben. Sie finden “... sie unter meiner Aufsicht in der Stadt bleiben würden ..." anderthalbmal so gut wie den Bezugssatz, und geben deshalb dem neuen Zitat 150 Punkte (= 100 mal 1,5).

Bei der Bewertung können Sie eine beliebige Brandbreite positiver Zahlen einschließlich Dezimalzahlen und Brüche verwenden. Sie können jedoch keine Null Punkte oder Minuszahlen vergeben. Versuchen Sie, eine möglichst große Bandbreite zu verwenden, damit Sie viele verschiedene Niveaus von “Richtigkeit" unterscheiden können.

Zur Zusammenfassung und Erinnerung

- Geben Sie dem Bezugssatz (Zitat 1) irgendeine Punktzahl, die Ihnen sinnvoll erscheint. (Diese wird am Anfang jeder Seite zur Erinnerung angegeben.)
- Bewerten Sie Zitate 2-30 in Bezug auf Zitat 1 (den Bezugssatz). Geben Sie korrekt klingenden Sätzen hohe Zahlen, und komisch oder unnatürlich klingenden Sätzen niedrige Zahlen.
- Versuchen Sie, so schnell wie möglich, aber ohne übermäßigen Zeitdruck eine Entscheidung zu treffen. Es gibt keine ‘richtigen’ oder ‘falschen’ Antworten - was zählt und uns interessiert, ist ihre spontane, unreflektierte Bewertung.

- Sie können eine beliebig große Bandbreite von Zahlen verwenden.
- Denken Sie daran, bei den Fragen danach, ob Sie den Film und den jeweiligen Ausschnitt kennen, immer auf ‘Ja’ oder ‘Nein’ zu klicken.
- Nehmen Sie als Bewertungsgrundlage ausschließlich den unterstrichenen Abschnitt des Gesamtzitats.
- Gehen Sie den Fragebogen der Reihe nach durch, ohne eine Seite zurückzuspringen oder die Seite neu zu laden.

Klicken Sie auf “Start!", um zum Film-Fragebogen zu kommen und mit der Studie zu beginnen.
[Start Button]

Instructions for Chinese Diathesis Experiment

关于电影剧本问卷的说明[3]

在这份问卷中，你将看到从电影剧本中摘取的30段对话。完成该问卷大约需要20分钟。你可能觉得有些句子完全没有问题，而另一些则不同。你的任务是决定一句话的中文有多地道。想象你正在帮助一个朋友学习中文，如果你的朋友使用同样的句子，听上去象是一个讲中文的人说的吗（不考虑发音因素）？还是听起来很别扭或不自然？

因为这些都是电影对话，所以风格很随意。你不必考虑一句话是否因为标点符号，俚语，或是讲话人在说话过程中出现犹疑、重复和更正等而显得不完美，例如："我要去...嗯...象...我父母家。"这个句子就不是问题，因为讲中文时经常会出现这种情况。此外，不要考虑你在学校学到的语法规则 — 答案没有对错，也请不要在一句话上花太多时间 — 你的第一印象才是我们感兴趣的。

如何给这些句子打分

你需要给每句话打分。你看到第一句话将作为"参考句"，而你将通过同这句话比较来为下面的句子打分。你可以选择任何你认为合适的分数来给参考句打分，然后在给下面的每个句子打分时，如果你认为一个句子比参考句更地道，就给更高的分数，如果你觉得句子听起来很奇怪或不自然，给的分就低一些。

例如，下面第一句话的中文还过得去，但可能有点别扭。让我们给它一个参考分100分。

1. 这个垫子是猫坐在上面的

你可能觉得第二句比第一句要好一倍，因为它听起来更自然，那么你可能给它打200分［即100乘以2］

2. 猫坐在垫子上

第三句话有明显的错误（"被坐着"而不是"坐在"）– 如果你觉得这句话比第一句差一倍，你可能给这句话50分［即100除以2］

3. 猫被坐着垫子上

最后一句话让人完全看不懂，所以如果你认为它相当于参考句的三倍那么差，你可以给这句话30分［大约为100分除以3］

4. 猫坐垫子上在

当然你可以采用任何你喜欢的参考分数。如果你给第1句打了20分，很可能就给第2句打了40分，第3句10分，而第4句则7分。

在你给每个句子打分时，我们还会给你一些背景情况：对电影的一小段描述及这段情景在电影中所处的部分。你应该先读一下这些背景描述，对你是否熟悉电影情节和问题中涉及的情景回答"是"或"否"（你不必对场景熟悉到逐字逐句的程度）。然后再读对话，对其中下划线部分打分。所以，对下面的问题范例，你应该仅仅基于对"...你来这里干什么..."的印象来打分。

[3] Original can be viewed at https://www.cs.tcd.ie/Brian.Murphy/bestestZH/instructions.html

［**电影**］ 少林前弟子星（周星驰）与他的少林师兄弟 和“黄金右腿”明锋（吴孟达）组成一支足球队来赢取中国足球杯。	“我看过这出电影” 是：[radio button] 否：[radio button]
［**情景**］ 阿梅（赵薇）作为守门员来挽救球赛。星觉得她的尼姑发型使她看上去像个外星人。	“我记得这段情景” 是：[radio button] 否：[radio button]
周星驰：你来这里干什么？ **赵薇**：我想帮你们比赛。 **周星驰**：...你快点回火星吧，地球是很危险地。	给划线部分打分： [textbox containing the number 150]

那么我们来一步步分析应该如何完成这个范例：

- 你在几年前看过“少林足球”，能依稀记得电影的情节，所以你对问题“我看过这出电影”回答“是”。
- 你不记得星取笑阿梅的发型这个情景，所以你对“我记得这段情景”这个问题回答“否”。
- 如果参考句是“你所有的护士把她排挤开”，而你给这句话打了100分。你认为“你来这里干什么？”相当于参考句一倍半那么好，所以你给它打150分。［100乘以1.5］

打分时，你可以使用任何你喜欢的正数，包括小数和分数，但不能使用0或负数，请尽量使用不同分数来将句子区分成尽可能多的不同水平。

请记住：

- 给参考句打任何你想给的分数（分数将在每页页首再次出现，以示提醒）。
- 给其它句子打分时，都参照这个分数，给更地道的句子打更高分，给听起来奇怪或不自然的句子更低分。.
- 快速但不要过于匆忙地给每个句子打分。答案没有对错 – 我们感兴趣的是你的第一印象。
- 可以将分数的差距随意拉大。
- 记得用“是”或“否”来告诉我们你是否熟悉这出电影和这段情景。
- 仅仅给对话的下划线部分打分。
- 请一直往下完成问卷，不要使用你浏览器上的“后退”或“重载”按键。

点击下键开始回答问卷。

Instructions for Judgement Method Experiment (Magnitude Estimation)

Instructions for Film Script Questionnaire[4]

In this questionnaire you will see 15 short extracts from film scripts - it should take you about 10-15 minutes to complete. Some of the quotes will seem perfectly OK to you, but others won't. Your task is to decide how English-sounding a particular quote is. Imagine you are helping a friend learn English. If your friend used the same sentence, would it sound like a native talking (ignoring pronunciation)? Or would it sound strange or unnatural?

Since these are film dialogues, the style is quite casual. You *don't* have to worry about whether a quote is bad because of punctuation, slang, hesitations, or repetitions and changes the speaker makes as they go along. For example, "I gotta go to the ... uh ... like ... my parents' place." would be fine, because English speakers often say things like that. Also, don't worry about any grammar rules you may have been told in school - there are no right or wrong answers. And don't spend too long on any quote - it is your *first impression* that we are interested in.

How to score the quotes

You will have to give each quote a score. The first quote you see will be the "reference quote", and you will score all other quotes compared to this. You can choose any number you want for the reference quote. Then, for each following quote, use a bigger number if you feel it is more English-sounding, and give a smaller number if you think it is strange or unnatural - there are no right or wrong answers.

For example, the first sentence below is okay English, but maybe a little awkward - let's give it a reference score of 100.

1. The mat was what the cat sat on

You might feel that the second sentence is twice as good, because it sounds much more natural. Then you might give it a score of 200. [that is 100 multiplied by two]

2. The cat sat on the mat

The third sentence has a clear mistake ('sitted' instead of 'sat') - if you felt it is twice as bad as sentence 1, you would give it a score of 50. [that is 100 divided by two]

3. The cat sitted on the mat

The last sentence is hard to make sense of at all, so if you find it three times worse than the reference you could give it a score of 30. [roughly, 100 divided by three]

4. cat sat on mat the

Of course you can use any reference score you want. If you had chosen 20 for sentence 1, you might have given 40 for sentence 2, 10 for sentence 3, and 7 for sentence 4.

When you are scoring a quote we will also give you a little bit of context - a short description of the film and the scene it belongs to. You should read these first and reply 'yes' or 'no' to the questions of whether you are *familiar* with the plots of the film and scene in question (you *don't* have to know the scene word for word). Then you should read the dialogue and give a score for the underlined portion of text. So, for the example question below, you should give a score based on your impression of "... they'd be left at the city ..." only.

[4] Original can be viewed at https://www.cs.tcd.ie/Brian.Murphy/bestestM/o/instructions.html

[Film] The Empire Strikes Back, George Lucas, 1980: The Star Wars rebel alliance continue to battle the evil Empire led by Darth Vader.	"I have seen this film" Yes: [radio button] No: [radio button]
[Scene] Darth Vader goes back on his deal with Lando Calrissian (Billy Dee Williams) to let Princess Leia (Carrie Fisher) and Chewbacca the Wookie go free.	"I remember this scene" Yes: [radio button] No: [radio button]
Darth Vader: ... Take the princess and the Wookie to my ship. **Lando:** You said they'd be left at the city under my supervision. **Darth Vader:** I am altering the deal. Pray I don't alter it any further.	Score for underlined text: [textbox containing the number 150]

So, to step through this example:

- You saw the film "The Empire Strikes Back" a long time ago, and you can roughly remember what it is about, so you answer "Yes" to the question "I have seen this film".
- You can't remember the scene where Leia and Chewbacca are taken prisoner, so you answer "No" to the question "I remember this scene".
- Say the reference quote was "I get them it for half", and it was given a score of 100. You think that "they'd be left at the city" is one and a half times as good as the reference quote, so you give it a score of 150. [100 multiplied by one and a half]

When scoring you can use any range of positive numbers that you like, including decimals and fractions. You can't use zero or negative numbers. Try to use a wide range of numbers to distinguish as many levels as possible.

Remember:

- Give the reference quote any number you want (it will be repeated at the start of each page as a reminder).
- Score all the other quotes relative to this one, giving big numbers for English-sounding ones and small numbers for strange or unnatural ones.
- Try to make up your mind as quickly as possible without rushing. There are no 'right' or 'wrong' answers - it is your first impression we are interested in.
- Use as wide a range of numbers as you want.
- Remember to mark 'yes' or 'no' to tell us if you are familiar with the scene and film.
- Base your score only on the underlined portion of the dialogue.
- Please go straight through the questionnaire, without using the 'back' or 'reload' buttons on your browser.

Click below to start the questionnaire.

[Start Button]

Instructions for Judgement Method Experiment (Likert Scale)

Instructions for Film Script Questionnaire[5]

In this questionnaire you will see 15 short extracts from film scripts - it should take you about 10-15 minutes to complete. Some of the quotes will seem perfectly OK to you, but others won't. Your task is to decide how English-sounding a particular quote is. Imagine you are helping a friend learn English. If your friend used the same sentence, would it sound like a native talking (ignoring pronunciation)? Or would it sound strange or unnatural?

Since these are film dialogues, the style is quite casual. You *don't* have to worry about whether a quote is bad because of punctuation, slang, hesitations, or repetitions and changes the speaker makes as they go along. For example, "I gotta go to the ... uh ... like ... my parents' place." would be fine, because English speakers often say things like that. Also, don't worry about any grammar rules you may have been told in school - there are no right or wrong answers. And don't spend too long on any quote - it is your *first impression* that we are interested in.

How to score the quotes

You will have to give each quote a score between 1 and 7. Use a bigger number if you feel it is more English-sounding, and give a smaller number if you think it is strange or unnatural - there are no right or wrong answers.

For example, the first sentence below is okay English, but maybe a little awkward - let's give it a reference score of 5.

1. The mat was what the cat sat on

You might feel that the second sentence is better, because it sounds much more natural. Then you might give it a score of 7.

2. The cat sat on the mat

The third sentence has a clear mistake ('sitted' instead of 'sat') - you might give it a score of 3 or 4.

3. The cat sitted on the mat

The last sentence is hard to make sense of at all, so it could be given a score of 1 or 2.

4. cat sat on mat the

When you are scoring a quote we will also give you a little bit of context - a short description of the film and the scene it belongs to. You should read these first and reply 'yes' or 'no' to the questions of whether you are *familiar* with the plots of the film and scene in question (you *don't* have to know the scene word for word). Then you should read the dialogue and give a score for the underlined portion of text. So, for the example question below, you should give a score based on your impression of "... they'd be left at the city ..." only.

[Film] The Empire Strikes Back, George Lucas, 1980: The Star Wars rebel alliance continue to battle the evil Empire led by Darth Vader.	"I have seen this film" Yes: [radio button] No: [radio button]
[Scene] Darth Vader goes back on his deal with Lando Calrissian (Billy Dee Williams) to let Princess Leia (Carrie Fisher) and Chewbacca the Wookie go free.	"I remember this scene" Yes: [radio button] No: [radio button]
Darth Vader: ... Take the princess and the Wookie to my ship. **Lando:** You said they'd be left at the city under my supervision. **Darth Vader:** I am altering the deal. Pray I don't alter it any further.	Score for underlined text: [drop-down menu with choice of numbers 1 to 7]

[5] Original can be viewed at https://www.cs.tcd.ie/Brian.Murphy/bestestM/l/instructionsLikert.html

So, to step through this example:

- You saw the film “The Empire Strikes Back" a long time ago, and you can roughly remember what it is about, so you answer “Yes" to the question “I have seen this film".
- You can’t remember the scene where Leia and Chewbacca are taken prisoner, so you answer “No" to the question “I remember this scene".
- You think that “they’d be left at the city" is perfect, so you give it a score of 7.

When scoring you can use any range of positive numbers that you like, including decimals and fractions. You can’t use zero or negative numbers. Try to use a wide range of numbers to distinguish as many levels as possible.

Remember:

- Give big numbers for English-sounding quotes and small numbers for strange or unnatural quotes.
- Try to make up your mind as quickly as possible without rushing. There are no ‘right’ or ‘wrong’ answers - it is your first impression we are interested in.
- Use as wide a range of numbers as you want.
- Remember to mark ‘yes’ or ‘no’ to tell us if you are familiar with the scene and film.
- Base your score only on the underlined portion of the dialogue.
- Please go straight through the questionnaire, without using the ‘back’ or ‘reload’ buttons on your browser.

Click below to start the questionnaire.

[Start Button]

Instructions for Judgement Method Experiment (Pairwise Comparison)

Instructions for Film Script Questionnaire[6]

In this questionnaire you will see 15 short extracts from film scripts - it should take you about 10-15 minutes to complete. Some of the quotes will seem perfectly OK to you, but others won't. Your task is to compare quotes, and decide which is more English-sounding. Imagine you are helping a friend learn English. If your friend used the same sentences, which would sound more like a native talking (ignoring pronunciation) and which would sound more strange or unnatural?

Since these are film dialogues, the style is quite casual. You *don't* have to worry about whether a quote is bad because of punctuation, slang, hesitations, or repetitions and changes the speaker makes as they go along. For example, "I gotta go to the ... uh ... like ... my parents' place." would be fine, because English speakers often say things like that. Also, don't worry about any grammar rules you may have been told in school - there are no right or wrong answers. And don't spend too long on any quote - it is your *first impression* that we are interested in.

How to judge the quotes

You will have to decide if you think each quote is better, the same, or worse compared to the one that comes before it - there are no right or wrong answers. A 'better' quote is one that you think is more English-sounding. A 'worse' quote is one that sounds strange or unnatural to you. If you think both sentences sound equally good or bad, or you can't decide, the quote is 'same'.

Let's look at some examples. The first sentence below is okay English, but maybe a little awkward.

1. The mat was what the cat sat on

If you think that the second sentence sounds more natural than the first sentence, you would label it 'better'.

2. The cat sat on the mat

The third sentence has a clear mistake ('sitted' instead of 'sat') - and so is probably 'worse' than sentence 2.

3. The cat sitted on the mat

The last sentence is hard to make sense of at all, so it is probably 'worse' again than sentence 3.

4. cat sat on mat the

Of course you can use any reference score you want. If you had chosen 20 for sentence 1, you might have given 40 for sentence 2, 10 for sentence 3, and 7 for sentence 4.

When you are deciding about a quote we will also give you a little bit of context - a short description of the film and the scene it belongs to. You should read these first and reply 'yes' or 'no' to the questions of whether you are *familiar* with the plots of the film and scene in question (you *don't* have to know the scene word for word). Then you should read the dialogue and decide whether the underlined portion of text is better or worse compared to the previous quote.

[The previous quote was "... I get them it for half ..."]

[6] Original can be viewed at https://www.cs.tcd.ie/Brian.Murphy/bestestM/p/instructionsPairwise.html

[Film] The Empire Strikes Back, George Lucas, 1980: The Star Wars rebel alliance continue to battle the evil Empire led by Darth Vader.	“I have seen this film" Yes: [radio button] No: [radio button]
[Scene] Darth Vader goes back on his deal with Lando Calrissian (Billy Dee Williams) to let Princess Leia (Carrie Fisher) and Chewbacca the Wookie go free.	“I remember this scene" Yes: [radio button] No: [radio button]
Darth Vader: ... Take the princess and the Wookie to my ship. **Lando:** You said they’d be left at the city under my supervision. **Darth Vader:** I am altering the deal. Pray I don’t alter it any further.	The underlined text is ... [drop-down menu with choice of “better”, “same” or “worse”] ... than the last one.

So, to step through this example:

- You saw the film “The Empire Strikes Back" a long time ago, and you can roughly remember what it is about, so you answer “Yes" to the question “I have seen this film".
- You can’t remember the scene where Leia and Chewbacca are taken prisoner, so you answer “No" to the question “I remember this scene".
- You think that “they’d be left at the city" is more natural sounding than the previous quote of “I get them it for half", so you label it ‘better’.

Remember:

- The previous quote will appear at the top of every page as a reminder.
- More English-sounding quotes are ‘better’ and, and more strange or unnatural quotes are ‘worse’. If they are the same, or you can’t decide, choose ‘same’.
- Try to make up your mind as quickly as possible without rushing. There are no ‘right’ or ‘wrong’ answers - it is your first impression we are interested in.
- Remember to mark ‘yes’ or ‘no’ to tell us if you are familiar with the scene and film.
- Base your judgement only on the underlined portion of the dialogue.
- Please go straight through the questionnaire, without using the ‘back’ or ‘reload’ buttons on your browser.

Click below to start the questionnaire.

[Start Button]

Instructions for Experiment on Effect of Context

Instructions for Film Script Questionnaire[7]

In this questionnaire you will see 30 short extracts from film scripts - it should take you about 20 minutes to complete. Some of the quotes will seem perfectly OK to you, but others won't. Your task is to decide how English-sounding a particular quote is. Imagine you are helping a friend learn English. If your friend used the same sentence, would it sound like a native talking (ignoring pronunciation)? Or would it sound strange or unnatural?

Since these are film dialogues, the style is quite casual. You *don't* have to worry about whether a quote is bad because of punctuation, slang, hesitations, or repetitions and changes the speaker makes as they go along. For example, "I gotta go to the ... uh ... like ... my parents' place." would be fine, because English speakers often say things like that. Also, don't worry about any grammar rules you may have been told in school - there are no right or wrong answers. And don't spend too long on any quote - it is your *first impression* that we are interested in.

How to score the quotes

You will have to give each quote a score. The first quote you see will be the "reference quote", and you will score all other quotes compared to this. You can choose any number you want for the reference quote. Then, for each following quote, use a bigger number if you feel it is more English-sounding, and give a smaller number if you think it is strange or unnatural.

For example, the first sentence below is okay English, but maybe a little awkward - let's give it a reference score of 100.

1. The mat was what the cat sat on

You might feel that the second sentence is twice as good, because it sounds much more natural. Then you might give it a score of 200. [that is 100 multiplied by two]

2. The cat sat on the mat

The third sentence has a clear mistake ('sitted' instead of 'sat') - if you felt it is twice as bad as sentence 1, you would give it a score of 50. [that is 100 divided by two]

3. The cat sitted on the mat

The last sentence is hard to make sense of at all, so if you find it three times worse than the reference you could give it a score of 30. [roughly, 100 divided by three]

4. cat sat on mat the

Of course you can use any reference score you want. If you had chosen 20 for sentence 1, you might have given 40 for sentence 2, 10 for sentence 3, and 7 for sentence 4.

When you are scoring a quote we will sometimes give you a little bit of context to help: the surrounding dialogue, a short description of the scene it belongs to, or maybe the film. You should read these first and reply 'yes' or 'no' to the questions of whether you are *familiar* with the plots of the film or scene in question (you *don't* have to know the scene word for word). Then you should give a score to the underlined portion of text. So, for the example question below, you should give a score based on your impression of "... they'd be left at the city ..." only.

[7] Original can be viewed at https://www.cs.tcd.ie/Brian.Murphy/bestestC/instructions.html

[Film] The Empire Strikes Back, George Lucas, 1980: The Star Wars rebel alliance continue to battle the evil Empire led by Darth Vader.	"I have seen this film" Yes: [radio button] No: [radio button]
[Scene] Darth Vader goes back on his deal with Lando Calrissian (Billy Dee Williams) to let Princess Leia (Carrie Fisher) and Chewbacca the Wookie go free.	"I remember this scene" Yes: [radio button] No: [radio button]
Darth Vader: ... Take the princess and the Wookie to my ship. **Lando:** You said they'd be left at the city under my supervision. **Darth Vader:** I am altering the deal. Pray I don't alter it any further.	Score for underlined text: [textbox containing the number 150]

So, to step through this example:

- You saw the film "The Empire Strikes Back" a long time ago, and you can roughly remember what it is about, so you answer "Yes" to the question "I have seen this film".
- You can't remember the scene where Leia and Chewbacca are taken prisoner, so you answer "No" to the question "I remember this scene".
- Say the reference quote was "I get them it for half", and it was given a score of 100. You think that "they'd be left at the city" is one and a half times as good as the reference quote, so you give it a score of 150. [100 multiplied by one and a half]

[*no film description*] "I think this film is": [empty textbox]	"I have seen this film" Yes: [radio button] No: [radio button]
[*no scene description*]	"I remember this scene" Yes: [radio button] No: [radio button]
Mayor Vaughn: Martin, it's all psychological. You yell "barracuda", everybody says, "Huh? What?". "Shark" is yelled by you, we've got a panic on our hands on the Fourth of July.	Score for underlined text: [empty textbox]

When scoring you can use any range of positive numbers that you like, including decimals and fractions. You can't use zero or negative numbers. Try to use a wide range of numbers to distinguish as many levels as possible.

Remember:

- Give the reference quote any number you want (it will be repeated at the start of each page as a reminder).
- Score all the other quotes relative to this one, giving big numbers for English-sounding ones and small numbers for strange or unnatural ones.
- Try to make up your mind as quickly as possible without rushing. There are no 'right' or 'wrong' answers - it is your first impression we are interested in.
- Use as wide a range of numbers as you want.

- Remember to mark ‘yes’ or ‘no’ to tell us if you are familiar with the scene and film; and try to guess the film title if it isn’t given.
- Base your score only on the underlined portion of the dialogue.
- Please go straight through the questionnaire, without using the ‘back’ or ‘reload’ buttons on your browser.

Click below to start the questionnaire.

[Start Button]

Instructions for Experiment on Effect of Context

Survey on Use of Thematic Roles[8]

Introduction

Thematic roles (or theta roles or semantic roles, or what you will) are widely appealed to in linguistics. There remain many disagreements on the details of what roles to include in an inventory, and exactly how to define them. However, their frequent use as a descriptive device, for example when discussing syntactic constructions like the passive or the dative forms, suggests that there may be broad agreement on their meaning.

We would like to test this. We have assembled an arbitrary set of sentences for linguists to annotate with role labels. We do not give any description or definition of labels, since people's existing understanding of them is part of what we are investigating.

If you are willing to participate, please make the annotation using the role labels listed below. If you introduce a label not included in this list, please make a note of why.

- Agent (aka Actor)
- Patient
- Experiencer
- Percept (aka Stimulus)
- Instrument
- Recipient
- Beneficiary
- Theme
- Manner
- Time
- Place
- Reason

For the moment we are not interested in the finer distinctions that can be made within roles, such as the agent/actor distinction (i.e. animate/volitional versus inanimate/non-volitional); different types of locational and temporal roles (e.g. source/goal/path/extent or duration/time point); or various kinds of contingencies (condition/reason/purpose/motivation etc).

We realise the risks of doing what could be seen as psycholinguistic research with linguists as informants. However, this exercise is primarily intended to identify 'best practice' in linguistic description.

What to do

For each of the text excerpts below (taken from film dialogues, or from websites), an immediate context is given. The clause under consideration for annotation is underlined, and the verb is italic. Following each excerpt, the relevant arguments are listed. For each of these choose the role label from the drop-down menu that you think best describes the relationship between verb and argument. You can use the text-box to the right of the drop-down menu for any notes you would like to add.

We realise that there are many sentences to annotate (81 in total). While we would like to get as many full responses as possible, we will of course accept and appreciate incomplete responses too. Think of it as way to relax over the holidays if you like!

An Example

[8] Original can be viewed at https://www.cs.tcd.ie/Brian.Murphy/survey/linguistRolesSurvey.html

In this example from Saturday Night Fever, the predicate in question is “spend". “I" can be considered as Agent, “a long time" as Time, and “my hair" might be annotated as Place, or possibly Patient.

[Film: Saturday Night Fever] **Tony Manero**: Would ya just watch the hair. Ya know, I spend a long time on my hair and he hit it; he hit my hair.

[drop down box with labels], Note: [text box for comment]: I
[drop down box with labels], Note: [text box for comment]: a long time
[drop down box with labels], Note: [text box for comment]: on my hair

When you are finished your annotation, please submit it by pressing the ‘Submit’ button at the bottom of the page. All replies received before the 30th of January will be included in a summary of results that I will post to the list.

Many Thanks and Regards,
Brian Murphy and Carl Vogel,
Computational Linguistics Group,
Trinity College Dublin.

Confidentiality

All raw data returned will remain confidential within the Computational Linguistics Group. Summary data will be included in published research. Email addresses or names will not be published or passed to any other party outside of the group.

However, we do intend to post a summary of the results to The Linguist List, and may include relevant comments from individual respondents. If you do not want to be individually identified by name and/or email in any summary posting to The Linguist List, please leave the name and/or address fields blank.

About You

* What is your name?
[text box] (optional)
* What is your email?
[text box] (optional)
* What variety of English do you speak?
[drop down box of dialects]
* Are you a native speaker of English?
Yes [radio button] or No [radio button]?

[experiment started here]

Appendix C

Additional Tables and Figures

C.1 Diathesis Experiment Results

Table C.1: English benefactive and dative sentences, ranked by success of diathesis

Original Sentence (Altered Sentence)	Diathesis	Anim.	Diff.	Orig.	IQR	Alt.	IQR
My mother's gonna make some fried peppers and sausage for us (My mother's gonna make us some fried peppers and sausage)	Benefactive	a	**0.16**	-0.16	0.55	0.00	0.58
fetch your slippers for you (fetch you your slippers)	Benefactive	a	**0.15**	-0.15	0.46	0.00	0.23
they couldn't design a survival suit for us (they couldn't design us a survival suit)	Benefactive	a	**0.00**	0.00	0.11	0.00	2.11
Remind me to send a thank you note to Mr. Boeing (Remind me to send Mr. Boeing a thank you note)	Dative	a	**0.00**	0.00	0.16	0.00	0.51
Is vengeance going to bring your son back to you (Is vengeance going to bring you your son back)	Dative	a	**0.00**	0.00	0.42	0.00	0.18
send a wire to the main office (send the main office a wire)	Dative	i	**0.00**	0.00	0.07	0.00	0.55
I'll catch this bird for you (I'll catch you this bird)	Benefactive	a	**-0.02**	-0.13	0.51	-0.15	0.62
you don't even send a dress to my house (you don't even send my house a dress)	Dative	i	**-0.13**	-0.16	0.52	-0.29	0.58
I sent the bill to your father (I sent your father the bill)	Dative	a	**-0.14**	0.14	0.66	0.00	0.12
get your supervisor for me (get me your supervisor)	Benefactive	a	**-0.15**	0.00	0.75	-0.15	0.32
I booked the pitch for Gryffindor (I booked Gryffindor the pitch)	Benefactive	i	**-0.16**	-0.04	0.63	-0.19	0.81
they chose a new form for him (they chose him a new form)	Benefactive	a	**-0.23**	-0.32	0.42	-0.55	0.55
How do you get funding for something like this (How do you get something like this funding)	Benefactive	i	**-0.26**	-0.07	0.33	-0.32	1.22
You will bring Captain Solo and the Wookie to me (You will bring me Captain Solo and the Wookie)	Dative	a	**-0.32**	0.00	0.37	-0.32	0.76
You may take Captain Solo to Jabba the Hutt (You may take Jabba the Hutt Captain Solo)	Dative	a	**-0.51**	0.00	0.20	-0.51	0.40
I should take my daughter to a witch doctor (I should take a witch doctor my daughter)	Dative	a	**-0.87**	0.00	0.15	-0.87	0.99

Anim: (a)nimacy or (i)nanimacy of the prepositional phrase referent in the original variant
Diff (difference score): difference between original and altered scores, corresponding to the success of the diathesis
Orig: median normalised magnitude estimation score given to original sentences (those having PP NP complements)
Alt: median normalised magnitude estimation score given to original sentences (those having NP NP complements)
IQR: Interquartile range of individual scores

Table C.2: English benefactive and dative sentences, ranked by success of diathesis (continued)

Original Sentence (Altered Sentence)	Diathesis	Anim.	Diff.	Orig.	IQR	Alt.	IQR
I'll find him for three (I'll find three him)	Benefactive	i	**-0.98**	-0.05	0.42	-1.03	3.32
Bring an Obscure Relative to Work (Bring Work an Obscure Relative)	Dative	i	**-1.00**	0.00	0.13	-1.00	0.74
my friend Sweet Jay took me to that video arcade in town (my friend Sweet Jay took that video arcade in town me)	Dative	i	**-1.17**	-0.15	0.32	-1.32	1.16
your years with us here at Rydell have prepared you for the challenges you face (your years with us here at Rydell have prepared the challenges you face you)	Benefactive	i	**-1.32**	0.00	0.14	-1.32	1.24
Send the fleet to the far side of Endor (Send the far side of Endor the fleet)	Dative	i	**-1.33**	0.17	0.37	-1.16	1.17
They've been making statues for some two thousand years (They've been making some two thousand years statues)	Benefactive	i	**-1.40**	-0.05	0.18	-1.45	1.11
It'll grind your bones for its bread (It'll grind its bread your bones)	Benefactive	i	**-1.42**	-0.32	0.85	-1.74	1.42
What in heaven's name brought you to Casablanca (What in heaven's name brought Casablanca you)	Dative	i	**-1.78**	0.32	0.58	-1.45	2.99

Table C.3: English related fillers sentences, ranked by success of diathesis

Original Sentence (Altered Sentence)	Diathesis	Anim.	Diff.	Orig.	IQR	Alt.	IQR
she was shunned by all you nurses (all you nurses shunned her)	Passive	a	**0.50**	-0.50	0.46	0.00	0.04
I'll get you water (I'll get water for you)	Benefactive	a	**0.09**	0.00	0.15	0.09	0.74
it is captured by the Nazis (the Nazis capture it)	Passive	i	**0.04**	-0.04	0.29	0.00	0.09
I didn't build you a stronger ship (I didn't build a stronger ship for you)	Benefactive	a	**0.00**	0.00	0.15	0.00	0.42
I'm granting you an audience (I'm granting an audience to you)	Dative	a	**0.00**	0.00	0.25	0.00	0.33
America owes you a debt of gratitude (America owes a debt of gratitude to you)	Dative	a	**0.00**	0.00	0.55	0.00	0.26
he was taken by these divers (these divers took him)	Passive	a	**0.00**	0.00	0.15	0.00	0.29
I've somehow been infected by it (it has somehow infected me)	Passive	a	**0.00**	0.00	0.21	0.00	0.00
that fire was caused by kooky wiring (kooky wiring caused that fire)	Passive	i	**-0.13**	-0.08	0.45	-0.21	0.32
I was overwhelmed by the power of this place (The power of this place overwhelmed me)	Passive	a	**-0.14**	0.14	0.50	0.00	0.15
It was made by those who are dead (Those who are dead made it)	Passive	i	**-0.25**	-0.16	0.58	-0.42	0.67
He'll bring them death (He'll bring death to them)	Dative	a	**-0.32**	0.00	0.07	-0.32	0.58
Its black gates are guarded by more than just orcs (More than just orcs guard its black gates)	Passive	i	**-0.32**	0.00	0.08	-0.32	0.36
you choose to find yourself another job (you choose to find another job for yourself)	Benefactive	a	**-0.42**	0.00	0.24	-0.42	0.19

Table C.5: English passive sentences, ranked by success of diathesis

Original Sentence (Altered Sentence)	Anim.	Diff.	Orig.	IQR	Alt.	IQR
Dinosaurs eat man (Man is eaten by dinosaurs)	a	**0.20**	-0.27	0.79	-0.08	0.42
my fingers touch brain (brain is touched by my fingers)	i	**0.19**	-0.71	0.94	-0.51	0.68
God wrought good men (Good men were wrought by God)	a	**0.11**	-0.26	1.00	-0.15	0.44
your father chose me (I was chosen by your father)	a	**0.00**	0.00	0.00	0.00	0.32
I loved her (she was loved by me)	a	**0.00**	0.00	0.20	0.00	0.33
The Empire will compensate you (You will be compensated by the Empire)	a	**0.00**	0.00	0.42	0.00	0.78
Our ships have sighted the Millennium Falcon (The Millennium Falcon has been sighted by our ships)	i	**0.00**	0.00	0.65	0.00	0.39
Someone saw a cockroach (A cockroach was seen by someone)	a	**-0.15**	0.00	0.20	-0.15	0.37
you salute the Captain (the Captain is saluted by you)	a	**-0.16**	0.00	0.02	-0.16	0.42
I understood the rules (the rules were understood by me)	i	**-0.17**	-0.15	0.42	-0.32	0.27
you served my father (my father was served by you)	a	**-0.37**	0.00	0.33	-0.37	0.87
He kissed me (I was kissed by him)	a	**-0.42**	0.00	0.42	-0.42	0.26
I hit him (He was hit by me)	a	**-0.42**	0.00	0.00	-0.42	0.19
I forbid it (it is forbidden by me)	i	**-0.42**	0.00	0.30	-0.42	0.65
you wasted it (it was wasted by you)	i	**-0.42**	0.00	0.04	-0.42	0.90
They cut the power (The power was cut by them)	i	**-0.42**	0.00	0.67	-0.42	0.58
You're saying this (This is being said by you)	i	**-0.42**	-0.10	0.16	-0.51	1.18
We've analyzed their attack (Their attack has been analyzed by us)	i	**-0.50**	0.00	0.29	-0.50	0.76
You seek the Oracle (The Oracle is sought by you)	a	**-0.51**	0.00	0.15	-0.51	0.58
I want 50% of ye plunder (50% of ye plunder is wanted by me)	i	**-0.58**	-0.29	0.26	-0.87	0.90
They're leaving Middle-earth (Middle-earth is being left by them)	i	**-0.66**	0.00	0.01	-0.66	0.79
You just flash that thing (That thing is just flashed by you)	i	**-0.70**	-0.11	0.92	-0.81	0.63
You know him (He is known by you)	a	**-0.71**	0.00	0.00	-0.71	1.01
I'm warning you (You're being warned by me)	a	**-0.71**	0.00	0.72	-0.71	1.29
you'll find the ones for this semester (the ones for this semester will be found by you)	i	**-0.87**	0.00	0.00	-0.87	0.35
I rang a few people (a few people were rung by me)	a	**-1.00**	0.00	0.42	-1.00	0.49
I hate the British (The British are hated by me)	a	**-1.00**	0.00	0.14	-1.00	0.50
you hit the wall (the wall is hit by you)	i	**-1.00**	0.00	1.42	-1.00	1.00
Crucifixion lasts hours (Hours are lasted by crucifixion)	i	**-1.17**	-0.15	0.43	-1.32	1.00
you hurry home (home is hurried by you)	i	**-1.58**	0.00	0.32	-1.58	1.00

Table C.6: German passive sentences, ranked by success of diathesis

Original Sentence (Altered Sentence)	Source	Anim.	Diff.	Orig.	IQR	Alt.	IQR
Dinosaurier fressen Adam (Adam wird von Dinosauriern gefressen) 'Dinosaurs eat man'	en	a	**0.19**	-0.19	0.51	0.00	0.19
Gott erschuf gute Menschen (Gute Menschen wurden von Gott erschaffen) 'God wrought good men'	en	a	**0.15**	-0.15	0.42	0.00	0.48
Nicht-Schwangere erschlägt wehrlosen Passanten mit Geburtstagstorte (Wehrloser Passant wird von Nicht-Schwangerer mit Geburtstagstorte erschlagen) 'not-pregnant woman kills defenceless pedestrian with birthday cake'	de	a	**0.00**	0.00	0.10	0.00	0.45
du hast dir die Chance vergeben (die Chance wurde von dir vergeben) 'you wasted it'	en	i	**0.00**	-0.74	1.05	-0.74	0.68
Unsere Schiffe haben den Millennium Falcon gesichtet (Der Millennium Falcon wurde von unseren Schiffen gesichtet) 'Our ships have sighted the Millennium Falcon'	en	i	**0.00**	0.00	0.24	0.00	0.24
dein Vater mich wählte (ich von deinem Vater gewählt wurde) 'your father chose me'	en	a	**-0.15**	-0.08	0.37	-0.23	0.51
Wir haben ihre Angriffsstruktur analysiert (Ihre Angriffsstruktur wurde von uns analysiert) 'We've analyzed their attack'	en	i	**-0.15**	0.00	0.15	-0.15	0.37
Unsere deutschen Waffenbrüder werden das schon gründlich machen (Das wird von unseren deutschen Waffenbrüdern schon gründlich gemacht werden) 'now our German brothers-in-arms will do it properly'	de	i	**-0.16**	0.00	0.19	-0.16	0.42
Jemand sah im 12. eine Kakerlake (Im 12. wurde von jemandem eine Kakerlake gesehen) 'Someone saw a cockroach'	en	a	**-0.17**	-0.15	0.74	-0.32	0.43
Er hat mich geküsst (Ich wurde von ihm geküsst) 'He kissed me'	en	a	**-0.23**	0.00	0.58	-0.23	0.76
hauen wir mal 'n paar Aale raus (werden mal 'n paar Aale von uns rausgehauen) 'we'll pull a couple of eels out'	de	a	**-0.31**	-0.23	0.51	-0.54	0.77
ich sie geliebt habe (sie von mir geliebt wurde) 'I loved her'	en	a	**-0.32**	0.00	0.13	-0.32	0.72
Die sehen uns nicht (Wir werden von denen nicht gesehen) 'they don't see us'	de	a	**-0.32**	0.00	0.00	-0.32	0.43
Hab ich dich beim Onanieren gestört (Bist du von mir beim Onanieren gestört worden) 'did I disturb you masterbating'	de	a	**-0.32**	0.00	0.16	-0.32	0.63

Table C.7: German passive sentences, ranked by success of diathesis (continued)

Original Sentence (Altered Sentence)	Source	Anim.	Diff.	Orig.	IQR	Alt.	IQR
Er flog uns direkt an (Wir wurden von ihm direkt angeflogen) 'he flew directly at us'	de	a	**-0.32**	0.00	0.37	-0.32	0.51
Das Imperium wird dich kompensieren (Du wirst durch das Imperium kompensiert werden) 'The Empire will compensate you'	en	a	**-0.32**	-0.42	1.58	-0.74	1.00
Nächstes Jahr finden Sie wohl die für jetzt (Nächstes Jahr werden wohl die für jetzt von Ihnen gefunden) 'you'll find the ones for this semester'	en	i	**-0.36**	-0.15	0.32	-0.51	0.68
Ich habe ihn geschlagen (Er wurde von mir geschlagen) 'I hit him'	en	a	**-0.42**	0.00	0.07	-0.42	0.96
Magic Maggy brauchte Spezialisten für einen gefährlichen Job (Für einen gefährlichen Job wurden von Magic Maggy Spezialisten gebraucht) 'Magic Maggy needed specialists for a dangerous job'	de	a	**-0.42**	0.00	0.14	-0.42	0.85
Filmen Sie einlaufende Besatzungen (Werden einlaufende Besatzungen von Ihnen gefilmt) 'you film crews going on board'	de	a	**-0.48**	-0.04	0.37	-0.51	0.68
habe ich ein paar Leute angerufen (wurden einige Leute von mir angerufen) 'I rang a few people'	en	a	**-0.51**	0.00	0.00	-0.51	0.68
Die haben den Strom abgestellt (Der Strom ist von denen abgestellt worden) 'They cut the power'	en	i	**-0.51**	0.00	0.47	-0.51	0.85
Sie verlassen Mittelerde (Mittelerde wird von ihnen verlassen) 'They're leaving Middle-earth'	en	i	**-0.51**	0.00	0.20	-0.51	0.63
Ich beneide sie um ihre Familie (Sie werden von mir um ihre Familie beneidet) 'I am envy her for her family'	de	a	**-0.56**	0.00	0.42	-0.56	0.89
ich Gehirn berühre (Gehirn von mir berührt wird) 'my fingers touch brain'	en	i	**-0.68**	-0.32	0.67	-1.00	1.58
Ich hasse die Briten (Die Briten werden von mir gehasst) 'I hate the British'	en	a	**-0.74**	0.00	0.00	-0.74	0.54
Du suchst das Orakel (Das Orakel wird von dir gesucht) 'You seek the Oracle'	en	a	**-0.79**	0.00	0.00	-0.79	0.84
der Arzt sagt, einen Tag noch ('einen Tag noch' wird vom Arzt gesagt) 'the doctor says, just one more day'	de	i	**-0.85**	-0.15	0.32	-1.00	1.00
Ich warne dich (Du wirst von mir gewarnt) 'I'm warning you'	en	a	**-0.91**	0.00	0.14	-0.91	0.91
ich verbiete es (es wird von mir verboten) 'I forbid it'	en	i	**-1.00**	0.00	0.37	-1.00	0.91
findest du schon irgendeinen anderen Deppenjob (wird schon irgendein anderer Deppenjob von dir gefunden) 'you'll find some other dumb job'	de	i	**-1.00**	0.00	0.32	-1.00	1.42

Table C.8: German passive sentences, ranked by success of diathesis (continued)

Original Sentence (Altered Sentence)	Source	Anim.	Diff.	Orig.	IQR	Alt.	IQR
brauchst du ein Fahrrad (wird ein Fahrrad von dir gebraucht) 'do you need a bike'	de	i	**-1.00**	0.00	0.07	-1.00	0.68
Ich verstehe überhaupt nichts mehr (Von mir wird überhaupt nichts mehr verstanden) 'I don't understand anything anymore'	de	i	**-1.00**	0.00	0.40	-1.00	1.27
sie die Wahrheit nie erfahren hat (die Wahrheit nie von ihr erfahren worden ist) 'she never learnt the truth'	de	i	**-1.00**	0.00	0.14	-1.00	0.32
du vorm Captain salutierst (vorm Captain von dir salutiert wird) 'you salute the Captain'	en	a	**-1.08**	0.00	0.15	-1.08	1.55
Sie kennen ihn (Er wird von Ihnen gekannt) 'You know him'	en	a	**-1.32**	0.00	0.17	-1.32	1.53
habt Ihr meinem Vater gedient (wurde meinem Vater von Euch gedient) 'you served my father'	en	a	**-1.32**	0.00	0.00	-1.32	1.32
Du blitzt einfach mit dem Ding (Mit dem Ding wird einfach von dir geblitzt) 'You just flash that thing'	en	i	**-1.32**	0.00	0.15	-1.32	0.98
Ich will 50% von Eurer Beute (50% von Eurer Beute werden von mir gewollt) 'I want 50% of ye plunder'	en	i	**-1.32**	0.00	0.15	-1.32	0.87
Du sagst das nur (Das wird nur von dir gesagt) 'You're saying this'	en	i	**-1.32**	0.00	0.32	-1.32	1.00
ihr in die Wall kommt (von euch in die Wall gekommen wird) 'you hit the wall'	en	i	**-1.32**	-1.00	1.76	-2.32	1.58
Ich brauch Shampoo (Shampoo wird von mir gebraucht) 'I need shampoo'	de	i	**-1.42**	0.00	0.34	-1.42	0.74
Hab ich es doch gesagt (Ist es doch von mir gesagt worden) 'there, just like I said'	de	i	**-1.46**	-0.07	0.32	-1.53	1.32
Sie heißt Walter (Walter wird sie geheißen) 'she is called Walter'	de	a	**-1.50**	0.00	0.15	-1.50	1.91
ich kannte die Vorschriften (die Vorschriften wurden von mir gekannt) 'I understood the rules'	en	i	**-1.74**	0.00	0.15	-1.74	1.32
kriegt jeder eine halbe Flasche Bier (wird von jedem eine halbe Flasche Bier gekriegt) 'everyone gets half a bottle of beer'	de	i	**-2.00**	0.00	0.15	-2.00	1.77
du bald nach Hause kommst (du bald nach Hause gekommen wirst) 'you hurry home'	en	i	**-2.25**	-0.07	0.32	-2.32	1.58
Die Kreuzigung dauert Stunden (Durch die Kreuzigung werden Stunden gedauert) 'Crucifixion lasts hours'	en	i	**-2.32**	0.00	0.08	-2.32	1.58
Die DDR wurde 40 (40 wurde durch die DDR geworden) 'The GDR was 40'	de	i	**-2.32**	0.00	0.22	-2.32	2.03
heißt das Kosmonaut (wird Kosmonaut dadurch geheißen) 'it is called cosmonaut'	de	a	**-2.74**	0.00	0.38	-2.74	1.66

Table C.9: German related filler sentences, ranked by success of diathesis

Original Sentence (Altered Sentence)	Source	Anim.	Diff.	Orig.	IQR	Alt.	IQR
wird die Zunge gekitzelt von drei Geschmäckern (kitzeln drei Geschmäcker die Zunge) 'the tongue is tickled by three tastes'	de	i	**0.55**	-0.57	0.96	-0.02	0.46
dass ich davon angesteckt wurde (dass sie mich angesteckt hat) 'I've somehow been infected by it'	en	a	**0.08**	-0.08	0.42	0.00	0.08
Er wurde von jenen angelegt, die tot sind (Jene, die tot sind, haben ihn angelegt) 'It was made by those who are dead'	en	i	**0.04**	-0.19	0.32	-0.15	0.32
Er ist von einem Taucher entführt worden (Ein Taucher hat ihn entführt) 'he was taken by these divers'	en	a	**0.00**	0.00	0.11	0.00	0.13
Ich wurde von der schöpferischen Kraft in diesem Park überwältigt (Die schöpferische Kraft in diesem Park überwältigte mich) 'I was overwhelmed by the power of this place'	en	a	**0.00**	-0.15	0.42	-0.15	0.42
wirst du vom Wasserdruck zerquetscht (zerquetscht dich der Wasserdruck) 'you'll be crushed by the water pressure'	de	a	**0.00**	0.00	0.15	0.00	0.08
er von den Nazis gefunden wird (die Nazis ihn finden) 'it is captured by the Nazis'	en	i	**0.00**	0.00	0.32	0.00	0.04
dieser Stuhl ist wohl von den jungen Spaniern vergessen worden (die jungen Spanier haben diesen Stuhl wohl vergessen) 'this stool here was left behind by the young Spaniards'	de	i	**0.00**	0.00	0.36	0.00	0.15
sie wurde auch von euch Schwestern gemieden (ihr Schwestern habt sie auch gemieden) 'she was shunned by all you nurses'	en	a	**-0.06**	-0.15	0.42	-0.22	0.84
Hanna ist beinahe von einem Perversen entführt worden (Ein Perverser hat Hanna beinahe entführt) 'Hanna was almost kidnapped by a pervert'	de	a	**-0.15**	0.00	0.07	-0.15	0.43
das Feuer durch Pfusch bei der Installation verursacht wurde (Pfusch bei der Installation das Feuer verursacht hat) 'that fire was caused by kooky wiring'	en	i	**-0.15**	0.00	0.15	-0.15	0.42
Seine schwarzen Tore werden von Schlimmerem als Orcs bewacht (Schlimmeres als Orcs bewachen seine schwarzen Tore) 'Its black gates are guarded by more than just orcs'	en	i	**-0.17**	-0.15	0.23	-0.32	0.74
Johannes Paul wurde auch hier von einer begeisterten Menge empfangen (Eine begeisterte Menge empfing auch hier Johannes Paul) 'here John Paul was also welcomed by an enthusiastic crowd'	de	a	**-0.23**	0.00	0.04	-0.23	0.63
Die meisten Kojen werden von 2 Mann benutzt (Die meisten Kojen benutzen zwei Mann) 'most of the bunks are used by 2 men'	de	i	**-0.55**	0.00	0.37	-0.55	0.68

Table C.10: Chinese *ba* sentences, ranked by success of diathesis

Original Sentence (Altered Sentence)	Source	Anim.	Diff.	Orig.	IQR	Alt.	IQR
你都浪费掉了时间 (你把时间都浪费掉了), 'you wasted it'	en	i	**0.80**	-0.74	0.21	0.06	0.78
它们切断了电力 (它们把电力切断了), 'They cut the power'	en	i	**0.74**	-0.15	0.61	0.58	0.37
你会找到今年的 (你会把今年的找到), 'you'll find the ones for this semester'	en	i	**0.15**	0.26	0.58	0.42	0.33
我们已分析了他们的攻击 (我们已把他们的攻击分析了。), 'We've analyzed their attack'	en	i	**0.00**	0.00	0.25	0.00	0.29
我要杀了你 (我要把你杀了), 'I will kill you'	en	a	**-0.17**	0.17	0.42	0.00	0.00
他有没有偷泡我的妞 (他有没有把我的妞偷泡), 'if he has swiped my girls'	zh	a	**-0.21**	0.42	0.38	0.21	1.71
你闪一下那个东西 (你把那个东西闪一下), 'You just flash that thing'	zh	i	**-0.21**	0.00	0.50	-0.21	0.67
我再遇上了他 (我再把他遇上了), 'I saw him again'	en	a	**-0.32**	0.00	0.64	-0.32	0.71
恐龙吃男人 (恐龙把男人吃), 'Dinosaurs eat man'	zh	a	**-0.35**	0.16	1.24	-0.19	0.00
他吻了我 (他把我吻了), 'He kissed me'	en	a	**-0.42**	0.42	0.00	0.00	0.10
我们的船舰看到千年鹰号了 (我们的船舰把千年鹰号看到了), 'Our ships have sighted the Millennium Falcon'	en	i	**-0.42**	-0.08	0.24	-0.50	1.29
我已经浪费了这一生 (我已经把这一生浪费了), 'I have already wasted this life'	en	i	**-0.51**	0.51	0.58	0.00	0.61
我曾经爱上一个人 (我曾经把一个人爱上), 'I once fell in love with somebody'	zh	a	**-0.58**	0.58	0.58	0.00	1.05
手指碰到脑浆 (手指把脑浆碰到), 'my fingers touch brain'	en	i	**-0.66**	0.00	1.00	-0.66	0.00
上帝创造了好人 (上帝把好人创造了), 'God wrought good men'	zh	a	**-0.75**	0.58	0.58	-0.16	0.66
我了解了游戏的规则 (我把游戏的规则了解了), 'I understood the rules'	en	i	**-0.81**	0.26	2.00	-0.55	0.00
我打了他 (我把他打了), 'I hit him'	en	a	**-0.94**	0.21	0.87	-0.74	0.66
你有没有碰过一种人 (你有没有把一种人碰过), 'have you ever met this kind of person'	en	a	**-0.95**	0.58	0.92	-0.37	1.00
他们要离开中土世界 (他们要把中土世界离开), 'They're leaving Middle-earth'	en	i	**-0.96**	0.00	0.21	-0.96	0.00
我竟然萌生了退出江湖的念头 (我竟然把退出江湖的念头萌生了), 'And now I'm thinking of quitting'	zh	i	**-0.97**	0.91	0.66	-0.07	1.00
你碰到水壁 (你把水壁碰到), 'you hit the wall'	en	i	**-1.00**	0.00	0.29	-1.00	1.00
你赶快回我们家 (你赶快把我们家回), 'you hurry home'	zh	i	**-1.00**	0.00	0.94	-1.00	1.06
我花十年 (我把十年花), 'I spent 10 years'	zh	i	**-1.00**	0.42	0.42	-0.58	0.50
我去了新加坡 (我把新加坡去了), 'I went to Singapore'	en	i	**-1.01**	0.14	0.42	-0.87	1.00
我看到你的牌了 (我把你的牌看到了), 'I can see your cards'	zh	i	**-1.01**	0.51	0.87	-0.50	2.16

Table C.11: Chinese *ba* sentences, ranked by success of diathesis (continued)

Original Sentence (Altered Sentence)	Source	Anim.	Diff.	Orig.	IQR	Alt.	IQR
帝国也会补偿你 (帝国也会把你补偿), 'The Empire will compensate you'	en	a	**-1.03**	0.45	0.79	-0.58	0.38
你来这里干什么 (你把这里来干什么), 'what did you come here for'	zh	i	**-1.05**	0.34	0.72	-0.71	0.00
我听到你的名字 (我把你的名字听到), 'When I hear your name'	zh	i	**-1.08**	0.42	0.79	-0.66	0.53
我打电话给认识他的人 (我把认识他的人打电话), 'I rang a few people'	en	a	**-1.11**	0.45	0.17	-0.66	0.32
我当拿你归案 (我当把你拿归案), 'I will bring you to justice'	zh	a	**-1.15**	0.42	0.58	-0.74	0.58
你有没有想过罐头的感受啊 (你有没有把罐头的感受想过啊), 'did you ever think about the tin's feelings?'	zh	i	**-1.19**	0.45	0.63	-0.74	1.04
都叫我一声"黎叔" (都把我一声叫"黎叔"), 'everyone calls me Uncle Li'	zh	a	**-1.28**	0.58	0.75	-0.69	1.79
你说这个 (你把这个说), 'You're saying this'	en	i	**-1.38**	0.58	0.32	-0.79	1.64
我认得你 (我把你认得), 'I know you'	zh	a	**-1.40**	0.66	0.78	-0.74	0.50
我问过老人们 (我把老人们问过), 'I asked the elders'	zh	a	**-1.42**	0.42	0.58	-1.00	0.29
某人看到了蟑螂 (某人把蟑螂看到了), 'Someone saw a cockroach'	en	a	**-1.53**	0.00	0.37	-1.53	1.79
你找先知的吗 (你把先知的找吗), 'You seek the Oracle'	en	a	**-1.58**	-0.42	0.00	-2.00	0.37
武当山收女弟子吗 (武当山把女弟子收吗), 'Does Wudang Mountain take female disciples?'	zh	a	**-1.66**	0.50	0.79	-1.16	2.21
你曾效命与我的父亲 (你曾把我的父亲效命), 'you served my father'	en	a	**-1.79**	0.00	1.21	-1.79	0.00
我这条命叫做「一将功成万骨枯」(我这条命把「一将功成万骨枯」叫做), 'I call this assignment "sacrifice for the greater good" '	zh	i	**-2.00**	0.00	1.32	-2.00	0.90
我痛恨英国人 (我把英国人痛恨), 'I hate the British'	zh	a	**-2.03**	0.66	0.76	-1.37	0.00
我回家等你 (我把家回等你), 'I'll go home and wait for you'	en	i	**-2.03**	0.00	0.30	-2.03	1.34
你父亲挑选了我 (你父亲把我挑选了), 'your father chose me'	en	a	**-2.16**	0.50	0.38	-1.66	1.53
我警告你 (我把你警告), 'I'm warning you'	en	a	**-2.16**	0.21	0.87	-1.95	0.00
你认识他吗 (你把他认识吗), 'You know him'	en	a	**-2.21**	0.42	0.71	-1.79	0.00
我很爱她 (我很把她爱), 'I loved her'	en	a	**-2.46**	0.14	0.42	-2.32	0.85
死刑可以延续好几个小时 (好几个小时可以被死刑延续), 'Crucifixion lasts hours'	en	i	**-2.74**	0.00	0.16	-2.74	1.37
我要你五成的战利品 (你五成的战利品被我要), 'I want 50% of ye plunder'	en	i	**-3.00**	1.00	0.00	-2.00	1.00

Table C.12: Chinese *bei* sentences, ranked by success of diathesis

Original Sentence (Altered Sentence)	Source	Anim.	Diff.	Orig.	IQR	Alt.	IQR
你都浪费掉了时间 (时间被你都浪费掉了), 'you wasted it'	en	i	**0.74**	-0.74	0.21	0.00	0.29
它们切断了电力 (电力被它们切断了), 'They cut the power'	en	i	**0.49**	-0.15	0.61	0.34	0.74
你来这里干什么 (这里被你来干什么), 'what did you come here for'	zh	i	**0.25**	0.34	0.72	0.58	0.00
我已经浪费了这一生 (这一生被我浪费了), 'I have already wasted this life'	zh	i	**0.11**	0.51	0.58	0.62	0.00
我们的船舰看到千年鹰号了 (千年鹰号被我们的船舰看到了), 'Our ships have sighted the Millennium Falcon'	en	i	**0.08**	-0.08	0.24	0.00	0.00
我打了他 (他被我打了), 'I hit him'	en	a	**0.06**	0.21	0.87	0.26	0.32
你碰到水壁 (水壁被你碰到), 'you hit the wall'	en	a	**0.00**	0.00	0.64	0.00	0.00
我再遇上了他 (他再被我遇上了), 'I saw him again'	zh	i	**0.00**	0.00	0.29	0.00	0.00
他吻了我 (我被他吻了), 'He kissed me'	en	a	**-0.04**	0.42	0.00	0.38	0.61
我看到你的牌了 (你的牌被我看到了), 'I can see your cards'	en	i	**-0.15**	0.51	0.87	0.37	0.00
恐龙吃男人 (男人被恐龙吃), 'Dinosaurs eat man'	zh	a	**-0.16**	0.16	1.24	0.00	0.09
上帝创造了好人 (好人被上帝创造了), 'God wrought good men'	en	a	**-0.25**	0.58	0.58	0.34	0.00
我了解了游戏的规则 (游戏的规则被我了解了), 'I understood the rules'	en	i	**-0.26**	0.26	2.00	0.00	0.55
你父亲挑选了我 (我被你父亲挑选了), 'your father chose me'	en	a	**-0.40**	0.50	0.38	0.10	0.38
我要杀了你 (你要被我杀了), 'I will kill you'	en	a	**-0.43**	0.17	0.42	-0.26	1.21
他们要离开中土世界 (中土世界要被他们离开), 'They're leaving Middle-earth'	zh	i	**-0.49**	0.00	0.21	-0.49	0.36
某人看到了蟑螂 (蟑螂被某人看到了), 'Someone saw a cockroach'	en	a	**-0.55**	0.00	0.37	-0.55	1.12
你找先知的吗 (先知的被你找吗), 'You seek the Oracle'	en	a	**-0.58**	-0.42	0.00	-1.00	1.16
我曾经爱上一个人 (一个人曾经被我爱上), 'I once fell in love with somebody'	en	a	**-0.66**	0.58	0.58	-0.08	0.00
手指碰到脑浆 (脑浆被手指碰到), 'my fingers touch brain'	en	i	**-0.79**	0.00	1.00	-0.79	0.00
死刑可以延续好几个小时 (好几个小时可以被死刑延续), 'Crucifixion lasts hours'	zh	i	**-0.79**	0.00	0.16	-0.79	1.05
你闪一下那个东西 (那个东西被你闪一下), 'You just flash that thing'	en	i	**-0.84**	0.00	0.50	-0.84	0.32
我们已分析了他们的攻击 (他们的攻击已分被我们分析了。), 'We've analyzed their attack'	en	i	**-1.00**	0.00	0.25	-1.00	0.90
你说这个 (这个被你说), 'You're saying this'	en	i	**-1.12**	0.58	0.32	-0.54	0.29
你有没有碰过一种人 (一种人有没有被你碰过), 'have you ever met this kind of person'	en	a	**-1.25**	0.58	0.92	-0.66	1.34
你会找到今年的 (今年的会被你找到), 'you'll find the ones for this semester'	zh	i	**-1.26**	0.26	0.58	-1.00	0.00

Table C.13: Chinese *bei* sentences, ranked by success of diathesis (continued)

Original Sentence (Altered Sentence)	Source	Anim.	Diff.	Orig.	IQR	Alt.	IQR
我很爱她 (她很被我爱), 'I loved her'	en	a	**-1.35**	0.14	0.42	-1.21	0.00
我认得你 (你被我认得), 'I know you'	zh	a	**-1.40**	0.66	0.78	-0.74	0.00
我竟然萌生了退出江湖的念头 (退出江湖的念头竟然被我萌生了), 'And now I'm thinking of quitting'	zh	i	**-1.41**	0.91	0.66	-0.50	1.29
我听到你的名字 (你的名字被我听到), 'When I hear your name'	zh	i	**-1.42**	0.42	0.79	-1.00	1.24
我回家等你 (家被我回等你), 'I'll go home and wait for you'	zh	i	**-1.50**	0.00	0.30	-1.50	1.16
我警告你 (你被我警告), 'I'm warning you'	en	a	**-1.53**	0.21	0.87	-1.32	1.81
我打电话给认识他的人 (认识他的人被我打电话), 'I rang a few people'	en	a	**-1.61**	0.45	0.17	-1.16	1.16
我去了新加坡 (新加坡被我去了), 'I went to Singapore'	zh	i	**-1.65**	0.14	0.42	-1.51	1.00
他有没有偷泡我的妞 (我的妞有没有被他偷泡), 'if he has swiped my girls'	zh	a	**-1.71**	0.42	0.38	-1.30	0.00
武当山收女弟子吗 (女弟子被武当山收吗), 'Does Wudang Mountain take female disciples?'	zh	a	**-1.74**	0.50	0.79	-1.24	0.00
你有没有想过罐头的感受啊 (罐头的感受有没有被你想过啊), 'did you ever think about the tin's feelings?'	zh	i	**-1.78**	0.45	0.63	-1.32	1.08
我问过老人们 (老人们被我问过), 'I asked the elders'	zh	a	**-1.78**	0.42	0.58	-1.37	1.21
你赶快回我们家 (我们家赶快被你回), 'you hurry home'	en	i	**-2.00**	0.00	0.94	-2.00	0.00
我痛恨英国人 (英国人被我痛恨), 'I hate the British'	en	a	**-2.03**	0.66	0.76	-1.37	1.58
你认识他吗 (他被你认识吗), 'You know him'	en	a	**-2.12**	0.42	0.71	-1.71	0.00
帝国也会补偿你 (你也会被帝国补偿), 'The Empire will compensate you'	en	a	**-2.27**	0.45	0.79	-1.82	0.00
我这条命叫做「一将功成万骨枯」(「一将功成万骨枯」被我这条命叫做), 'I call this assignment "sacrifice for the greater good" '	zh	i	**-2.32**	0.00	1.32	-2.32	0.00
我当拿你归案 (你被我当拿归案), 'I will bring you to justice'	zh	a	**-2.42**	0.42	0.58	-2.00	0.00
你曾效命与我的父亲 (我的父亲被你曾效命), 'you served my father'	en	a	**-2.45**	0.00	1.21	-2.45	2.11
我要你五成的战利品 (你五成的战利品被我要), 'I want 50% of ye plunder'	en	i	**-2.63**	1.00	0.00	-1.63	0.00
我花十年 (我把十年花), 'I spent 10 years'	zh	i	**-3.15**	0.42	0.42	-2.74	0.00
都叫我一声"黎叔" (都把我一声叫"黎叔"), 'everyone calls me Uncle Li'	zh	a	**-3.17**	0.58	0.75	-2.58	0.00

Table C.14: Chinese related fillers, ranked by success of diathesis

Original Sentence (Altered Sentence)	Diathesis	Source	Anim.	Diff.	Orig.	IQR	Alt.	IQR
它被纳粹夺走 'it is captured by the Nazis'	*bei*	en	i	**0.35**	0.45	0.86	0.79	0.50
半兽人把黑色大门看守 'Its black gates are guarded by more than just orcs'	*ba*	en	i	**0.32**	-0.32	1.39	0.00	0.50
那场火被老旧的管线引起了 'that fire was caused by kooky wiring'	*bei*	en	i	**0.32**	-0.32	1.32	0.00	0.07
各自把一张牌放在对方口袋里 'each of them puts one card in eachother's pocket'	*ba*	zh	i	**0.28**	0.17	0.50	0.45	0.41
她把孩子打掉啦 'she got rid of the child'	*ba*	zh	a	**0.17**	0.42	0.72	0.58	0.42
你把我感染了 'I've somehow been infected by it'	*ba*	en	a	**0.15**	-0.15	0.87	0.00	0.77
他被一个潜水人带走了 'he was taken by these divers'	*bei*	en	a	**-0.08**	0.00	0.46	-0.08	0.00
亡灵把天罗地网撒下 'It was made by those who are dead'	*ba*	en	i	**-0.08**	0.22	0.58	0.13	0.00
我把你衣服弄破了 'I damaged your clothes'	*ba*	zh	i	**-0.45**	0.89	0.49	0.45	0.86
我竟然被两个叫阿May的人抛弃。 'I would be brushed off by two people called May'	*bei*	zh	a	**-0.50**	0.50	0.57	0.00	1.58
俺把眼珠子抠出来 'I'll gorge my eye out'	*ba*	zh	i	**-0.58**	0.42	0.24	-0.16	0.00
电被人断了。 'the electricity has been cut off'	*bei*	zh	i	**-0.78**	0.42	0.29	-0.37	0.87
我曾被这里的力量所感动 'I was overwhelmed by the power of this place'	*bei*	en	a	**-1.42**	0.00	0.37	-1.42	2.58
露露在她房里被男朋友插了几刀 'In her room Meimei had been stabbed several times by her boyfriend'	*bei*	zh	a	**-2.00**	0.42	0.87	-1.58	0.79

C.2 Data Histograms

Some parametric statistical test demand that the sampling distribution of the data is normal (Gaussian). If the underlying data distribution is Gaussian, by definition its sampling distribution is also. Otherwise LaPlace's Central Limit Theory states that the sampling distribution of non-Gaussian data will tend towards normality as the sample size increases – the more skewed the data is, the larger a sample is necessary to approach normality.

By its nature, linguistic acceptability data is strongly skewed to the negative side. While participants may find some sentences radically worse than an average authentic utterance, they almost never will find a sentence that is radically better than average.

For example, the median acceptability score among all 155 variants of all sentences in the English acceptability experiment (see sections 3.2.2, 3.2.3) was 0 (i.e. no difference from the reference). The majority of original sentences received this aggregate score – an example of one is (C.1). There were only seven sentences whose median aggregate scores were higher than this – the highest scoring (C.2) came in at 0.32 (i.e. is was rated as about 25% better than the reference). On the other hand there were 78 sentences that fared worse, to a minimum score of -1.74 (i.e. three times worse than the reference), such as (C.3).

(C.1) ... I didn't build you a stronger ship.[1]

(C.2) What in heaven's name brought you to Casablanca?[2]

(C.3) Stop thinking sex about![3]

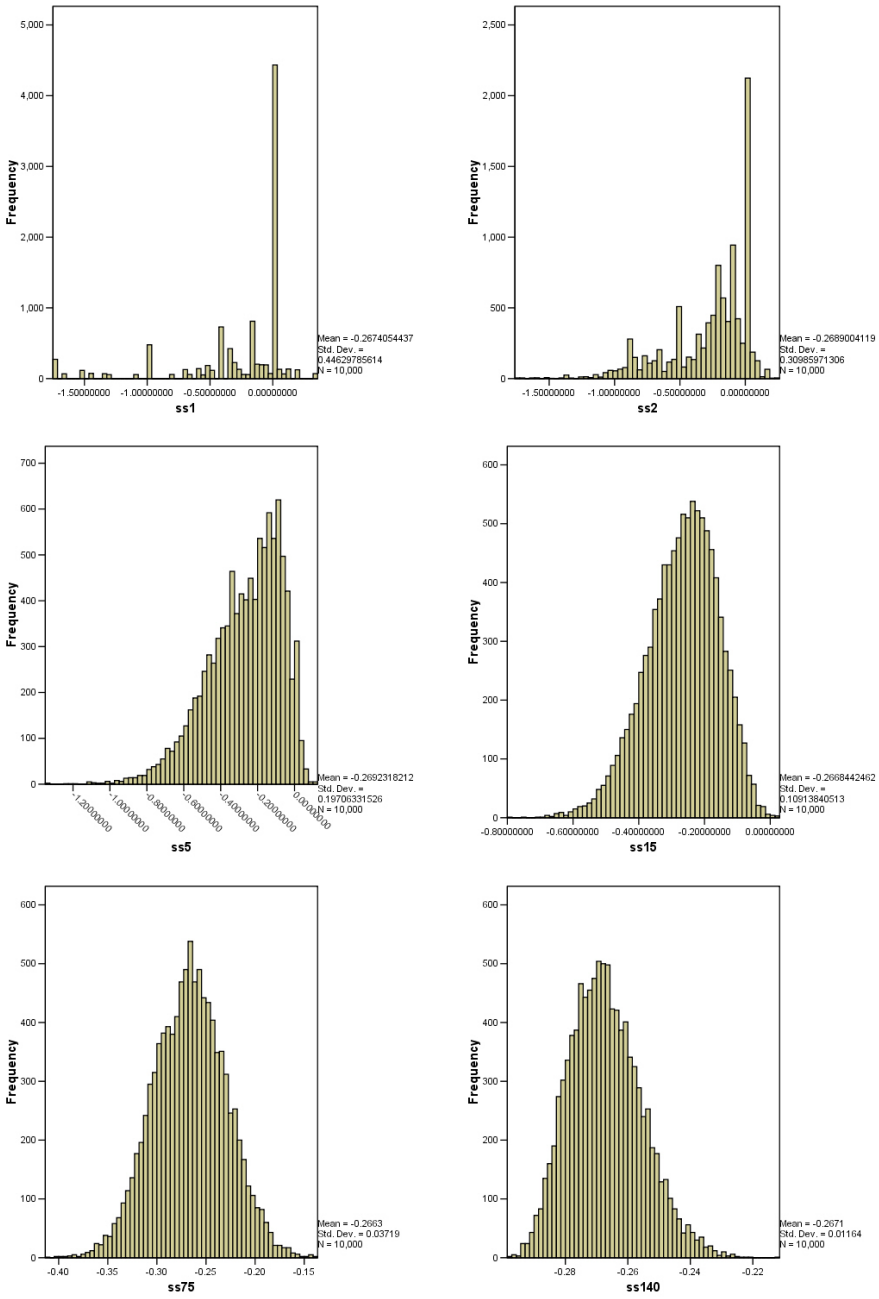
This results is the distribution in the top-left panel (marked "ss1") of figure C.1, that has a strong core of average sentences, a very small right tail of unusually good sentences, and a long wide left tail of increasingly bad sentences.

The histograms here are calculated on the basis of simulated sampling distributions. For sample size of interest (minimum: 1; maximum: the size of the data set), ten thousand random sub-samples were taken, and a mean was calculated. For each sample size, the shape of the histogram gives an estimated indication of the normality of its sampling distribution.

Figure C.1 suggests that for the relatively heterogenous data set of all sentences in the English diathesis experiment, a sample-size of 5 is inadequate, yielding a strongly skewed sampling distribution. However, a sample size of 15 is little different from one of 75, both having a strong central bell curve, and only marginal skewing to the negative tail.

[1] Film: *Titanic*, James Cameron, 1997.
[2] Film: *Casablanca*, Michael Curtiz, 1942.
[3] Film: *Life of Brian*, Monty Python, 1979.

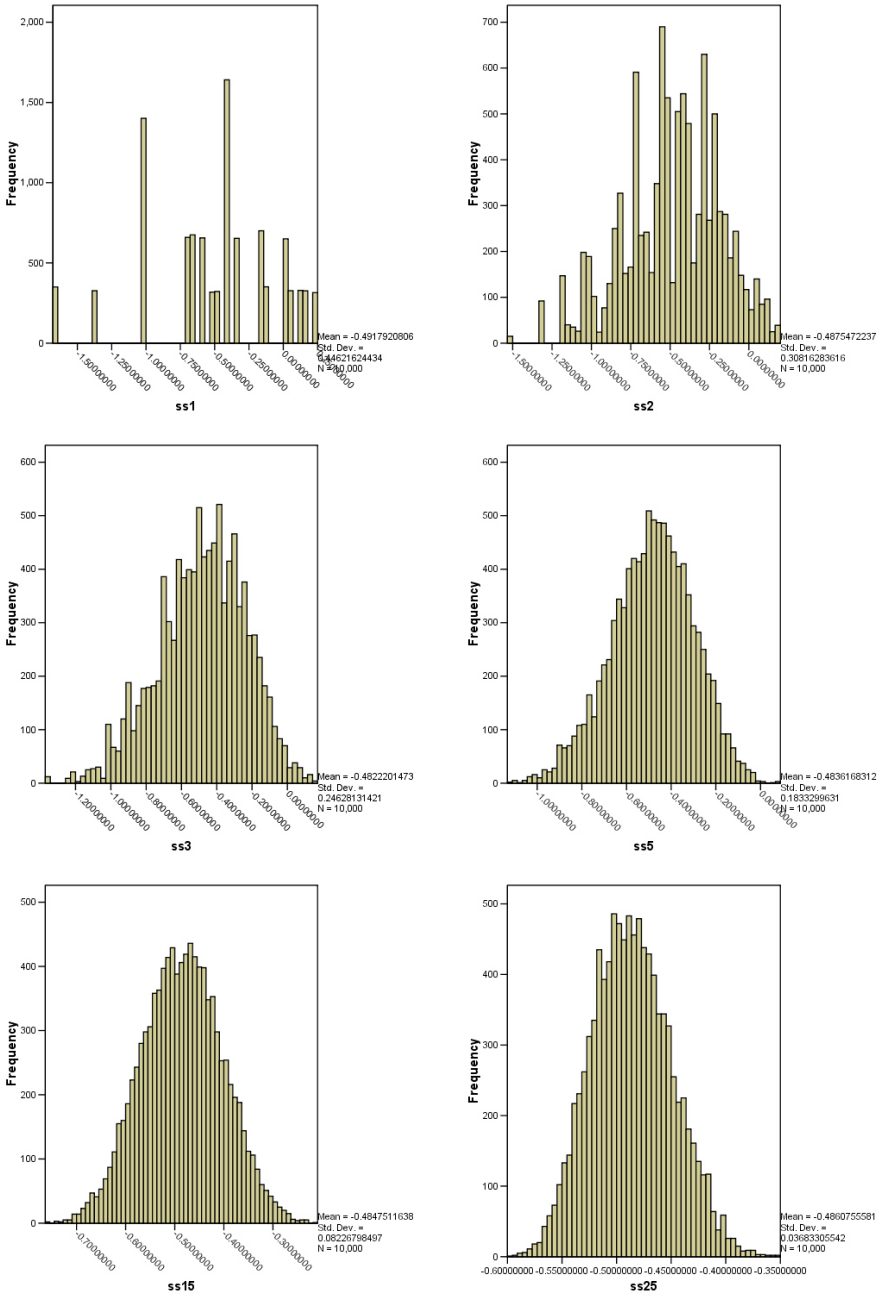
Figure C.1: Histograms of simulated sampling distributions on acceptability of English experimental sentences



Histograms of simulated sampling distributions on acceptability of all 155 English sentence items (including both original and altered variants).
X-axis is sample mean of measures of acceptability in normalised magnitude estimation scores.
Sample size of X is indicated "ssX".

The more homogenous data of the passive/active subset (figure C.2) does not need such large samples to approximate a Gaussian distribution – a sample size of 5 seems adequate.

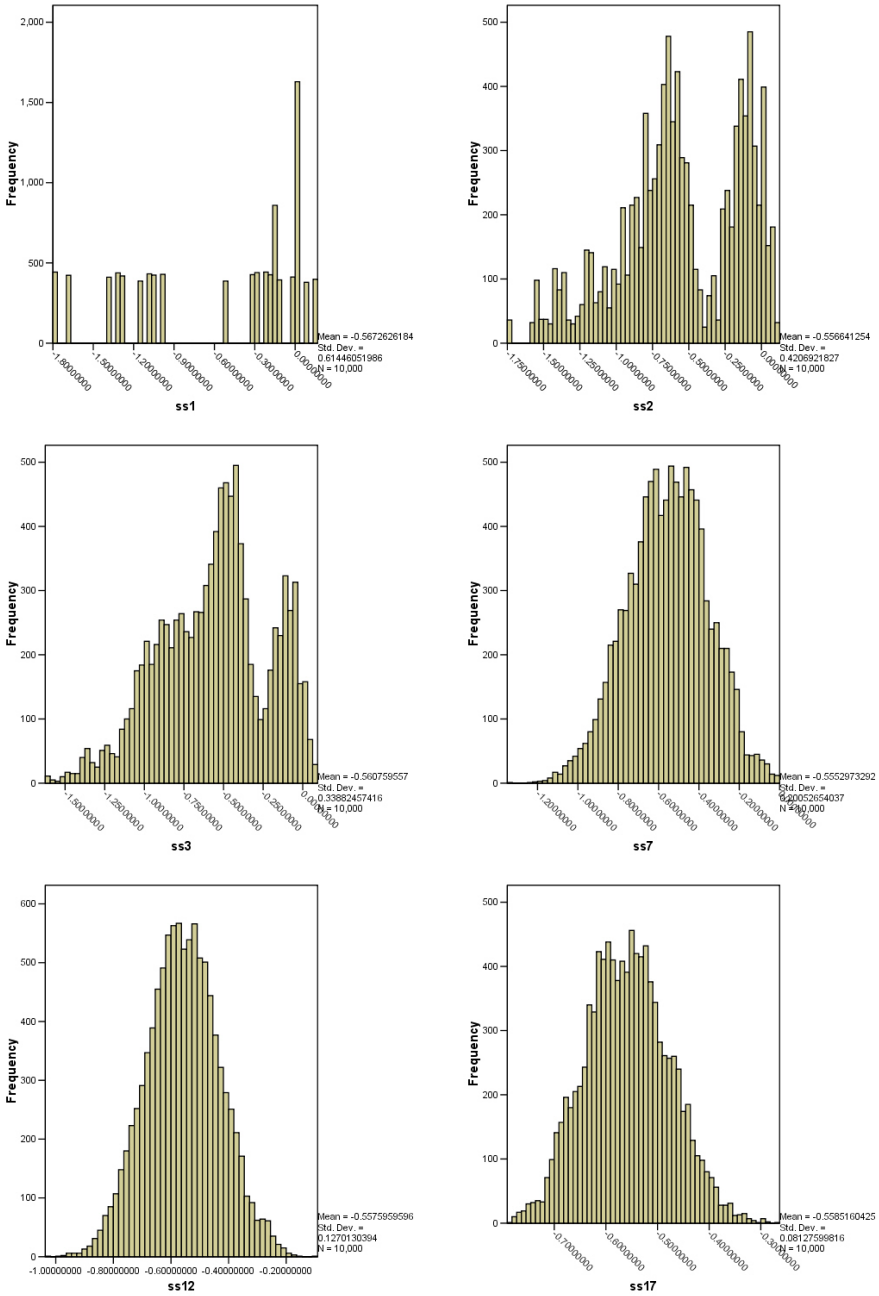
Figure C.2: Histograms of simulated sampling distributions on acceptability of English experimental passive diatheses



Histograms of simulated sampling distributions on relative acceptability between 30 English active/passive variant pairs.
X-axis is sample mean of measures of acceptability in normalised magnitude estimation scores.
Sample size of X is indicated "ssX".

The dative/benefactive data (figure C.3) is relatively homogenous but is bipolar, and so needs a sample of at least 7 to approximate a Gaussian distribution.

Figure C.3: Histograms of simulated sampling distributions on acceptability of English experimental dative/benefactive diatheses

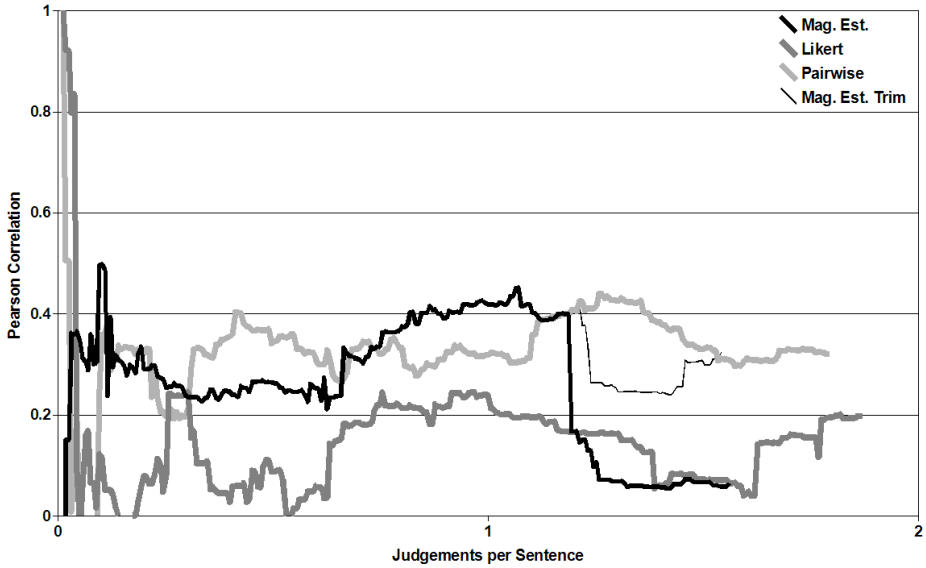


Histograms of simulated sampling distributions on relative acceptability between 24 dative/benefactive variant pairs.
X-axis is sample mean of measures of acceptability in normalised magnitude estimation scores.
Sample size of X is indicated "ssX".

C.3 Convergence of Rating Scale Data

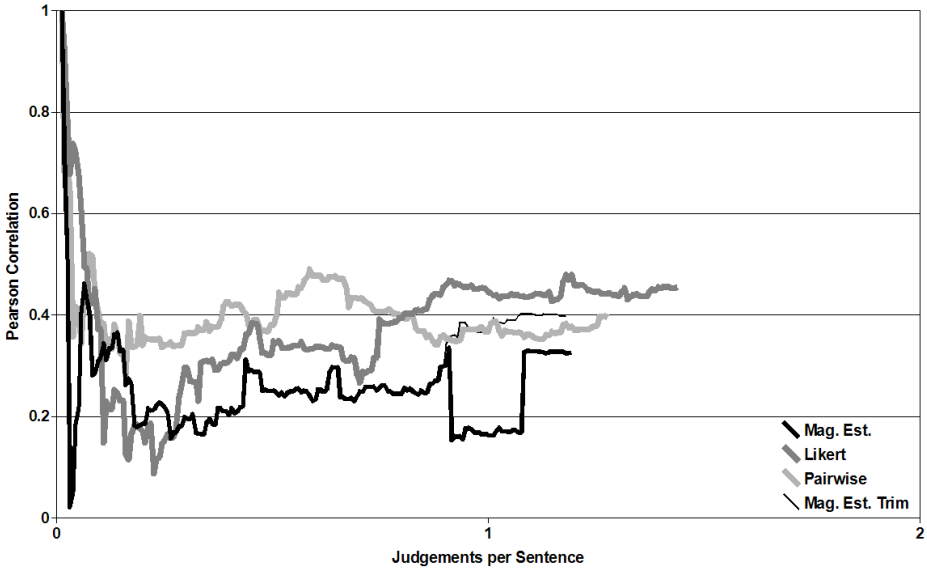
These graphs show the development of pairwise correlations between the (magnitude estimate) benchmark English acceptability data and the data from each of three judgement scale subexperiments. Means are used as aggregate measure of central tendency for individual stimuli.

Figure C.4: Correlation of data from three measurement scales to a magnitude estimation benchmark; upper tercile of sentences (most acceptable)



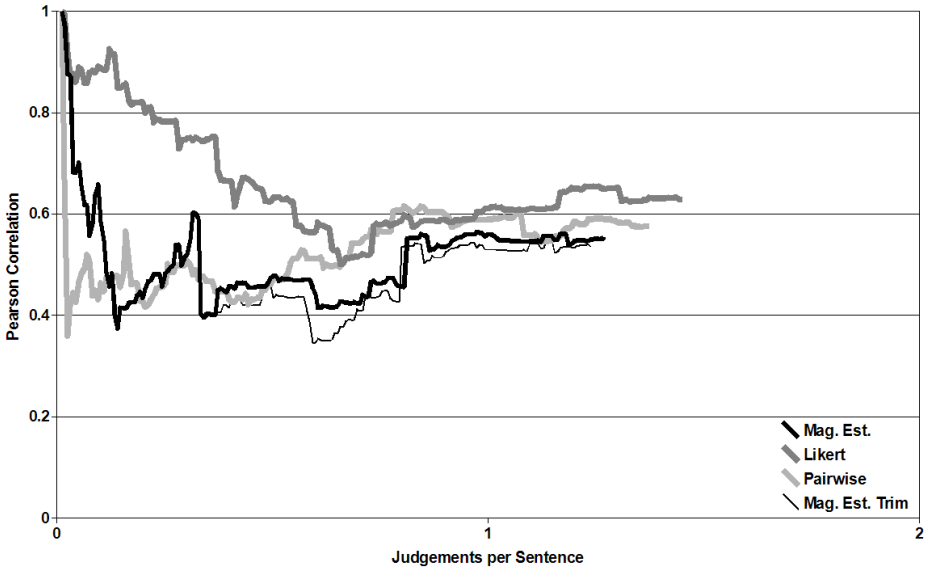
Pearson (linear) correlation over acceptability stimuli between data gathered using three rating scales.
Trimmed magnitude estimation data set contains only responses within range of plus/minus three standard deviations.

Figure C.5: Correlation of data from three measurement scales to a magnitude estimation benchmark; middle tercile of sentences (marginally acceptable)



Pearson (linear) correlation over acceptability stimuli between data gathered using three rating scales.
Trimmed magnitude estimation data set contains only responses within range of plus/minus three standard deviations.

Figure C.6: Correlation of data from three measurement scales to a magnitude estimation benchmark; lower tercile of sentences (least acceptable)



Pearson (linear) correlation over acceptability stimuli between data gathered using three rating scales.
Trimmed magnitude estimation data set contains only responses within range of plus/minus three standard deviations.

Bibliography

Aissen, Judith. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory* 21, 435–483.

Anderson, John. 2004. Modern grammars of case: a personal history. Lectures, Université de Toulouse-Le Mirail.

Anderson, Stephen R. 1971. On the Role of Deep Structure in Semantic Interpretation. *Foundations of Language* 7, 387–396.

Aristotle. 4C BC. *Physica*. Loeb classical library, London: Heinemann, with an English translation by Philip H. Wicksteed and Francis M. Cornford (1934).

Arnold, Jennifer E., Losongco, Anthony, Wasow, Thomas and Ginstrom, Ryan. 2000. Heaviness vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering. *Language* 76(1), 28–55.

Arregi, Karlos. 2001. Focus and Word Order in Basque. Manuscript, available from homepage: https://netfiles.uiuc.edu/karlos/www/.

Artstein, Ron and Poesio, Massimo. 2007. Inter-Coder Agreement for Computational Linguistics, submitted manuscript.

Baayen, R. Harald, Davidson, Doug J. and Bates, Douglas M. 2007. Mixed effects. modelling with crossed random effects for subjects and items. manuscript.

Baker, Mark. 1997. Thematic roles and syntactic structure. In Haegeman Liliane (ed.), *Elements of Grammar*, pages 73–137, Dordrecht: Kluwer.

Baker, Mona. 1992. *In other words: a coursebook on translation*. London: Routledge.

Bard, Ellen, Robertson, Dan and Sorace, Antonella. 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1), 32–68.

Barker, Chris and Dowty, David. 1993. Non-verbal Thematic Proto-Roles. In Amy Schafer (ed.), *Proceedings of NELS 23, Vol. 1*, pages 49–62, Amherst: GSLA.

Barnwell, Katharine. 1974. *Introduction to semantics and translation: with special reference to Bible translation*. High Wycombe: Summer Institute of Linguistics.

Beavers, John. 2004. Semantic Underspecificity in English Argument/Oblique Alternations. In *Proceedings of WECOL 2004*, Los Angeles: University of Southern California.

Bender, Emily. 2000. The Syntax of Mandarin Ba: Reconsidering the Verbal Analysis. *Journal of East Asian Linguistics* 9, 105–145.

Berger, Peter L. and Luckmann, Thomas. 1967. *The social construction of reality: a treatise in the sociology of knowledge*. Harmondsworth: Penguin.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan and Finegan, Edward. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Bisang, Walter. 2006. From meaning to syntax – semantic roles and beyond. In Ina Bornkessel, Matthias Schlesewsky, Bernard Comrie and Angela D. Friederici (eds.), *Semantic Role Universals and Argument Linking: Theoretical, Typological and Psycholinguistic Perspectives*, pages 191–236, Mouton de Gruyter.

Blatt, A., Luckhardt, Heinz-Dirk and Truar, M. 1984. Evaluation of Linguistic Specifications LS.1 Concerning Semantic Relations. Technical Report, European Community, DGDXIII.

Bock, Kathryn, Butterfield, Sally, Cutler, Anne, Cutting, J. Cooper, Eberhard, Kathleen M. and Humphreys, Karin R. 2006. Number Agreement in British and American English: Disagreeing to Agree Collectively. *Language* 82(1), 64–113.

Bock, Kathryn, Eberhard, Kathleen M. and Cutting, J. Cooper. 2004. Producing number agreement: How pronouns equal verbs. *Journal of Memory and Language* 51(2), 251–278.

Bock, Kathryn and Levelt, Willem. 1994. Language Production: Grammatical Encoding. In *Handbook of Psycholinguistics*, London: Academic Press.

Bock, Kathryn, Loebell, Helga and Morey, Randal. 1992. From conceptual roles to structural relations: bridging the syntactic cleft. *Psychological Review* 99, 150–171.

Boersma, Paul. 2004. A Stochastic OT account of paralinguistic tasks such as grammaticality and prototypicality judgments, unpublished manuscript.

Bohas, Georges, Guillaume, Jean-Patrick and Kouloughli, Djamel Eddine. 1990. *The Arabic linguistic tradition*. London: Routledge.

Bornkessel, Ina and Schlesewsky, Martin. 2006. Generalised semantic roles and syntactic templates: A new framework for language comprehension. In Ina Bornkessel, Matthias Schlesewsky, Bernard Comrie and Angela D. Friederici (eds.), *Semantic Role Universals and Argument Linking: Theoretical, Typological and Psycholinguistic Perspectives*, pages 327–353, Mouton de Gruyter.

Brennan, Eileen M., Ryan, Ellen B. and Dawson, William E. 1975. Scaling of apparent accentedness by magnitude estimation and sensory modality matching. *Journal of Psycholinguistic Research* 4(1), 27–36.

Bresnan, Joan, Cueni, Anna, Nikitina, Tatiana and Baayen, Harald. 2007. Predicting the Dative Alternation. In G. Boume, I. Kraemer and J. Zwarts (eds.), *Cognitive Foundations of Interpretation*, pages 69–94, Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan, Dingare, Shipra and Manning, Chris. 2001. Soft constraints mirror hard constraints: voice and person in English and Lummi. In *Proceedings of the LFG01 Conference*, University of Hong Kong.

Bresnan, Joan and Nikitina, Tatiana. 2003. On the Gradience of the Dative Alternation, unpublished manuscript.

Brill, Eric. 1992. A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento.

Burke, Michael, Cahill, Aoife, McCarthy, Mairead, O'Donovan, Ruth, van Genabith, Josef and Way, Andy. 2004. Evaluating Automatic LFG F-Structure Annotation for the Penn-II Treebank. *Journal of Research on Language and Computation; Special Issue on Treebanks and Linguistic Theories* .

Burnage, Gavin and Dunlop, Dominic. 1992. Encoding the British National Corpus. In Jan Aarts, Pieter de Haan and Nelleke Oostdijk (eds.), *English Language Corpora: Design, Analysis and Exploitation: Papers from the 13th international conference on English Language research on computerized corpora*, Amsterdam: Rodopi Press.

Burnard, Lou. 1995. *Users Reference Guide, British National Corpus, Version 1.0*. Oxford University Computing Services/British National Corpus Consortium, Oxford.

Carletta, Jean C. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics* 22(2), 249–254.

Carnie, Andrew. 1995. Complex Predicates and Deriving Copular Word Order. In *Proceedings of the Language in Ireland Conference, Belfast Working Papers in Linguistics, Vol 13*, pages 21–46, Jordanstown.

Chang, Franklin, Bock, Kathryn and Goldberg, Adele E. 2003. Can thematic roles leave traces in their places? *Cognition* 90, 29–49.

Charniak, Eugene. 2000. A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL-2000*.

Chen Ping. 2004. Identifiability and definiteness in Chinese. *Linguistics* 42(6), 1129–1184.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, Noam. 1958. A Transformational Approach to Syntax. In Jerry A. Fedor and Jerrold J. Katz. (eds.), *The Structure of Language: readings in the philosophy of language*, Chapter 7, pages 211–245, New York: Prentice-Hall.

Chomsky, Noam. 1964. Degrees of grammaticalness. In Jerry A. Fodor and Jerrold J. Katz (eds.), *The structure of language: readings in the philosophy of language*, pages 385–389, Englewood Cliffs: Prentice-Hall.

Chomsky, Noam. 1981. *Lectures on government and binding: the Pisa lectures*. Dordrecht: Foris.

Chomsky, Noam. 1986. *Knowledge of language: its nature, origin, and use*. New York: Praeger.

Chomsky, Noam. 1995. *The minimalist program*. Cambridge: MIT Press.

CKIP Group. 1998. An Introduction to the Academia Sinica Balanced Corpus of Modern Chinese. Technical Report, Chinese Knowledge Information Processing Group, Institute of Information Science, Academia Sinica, Taipei.

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph. D.thesis, University of Pennsylvania.

Comrie, Bernard and van den Berg, Helma E. 2006. Experiencer constructions in Daghestanian languages. In Ina Bornkessel, Matthias Schlesewsky, Bernard Comrie and Angela D. Friederici (eds.), *Semantic Role Universals and Argument Linking: Theoretical, Typological, and Psycholinguistic Perspectives*, pages 127–154, Berlin: Mouton de Gruyter.

Cowart, Wayne. 1997. *Experimental syntax*. London: Sage Publications.

Croft, William. 1991. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information.*. Chicago: University of Chicago Press.

Croft, William. 1993. Case marking and the semantics of mental verbs. In James Pustejovsky (ed.), *Semantics and the Lexicon*, Chapter 5, pages 55–72, Dordrecht: Kluwer Academic.

Croft, William. 2007. The origins of grammar in the verbalization of experience. *Cognitive Linguistics* 18(3), 339–382.

Cunningham, Hamish, Maynard, Diana, Bontcheva, Kalina and Tablan, Valentin. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia.

Davis, Anthony. 1996. *Lexical Semantics and Linking in the Hierarchical Lexicon*. Ph. D.thesis, Department of Linguistics, Stanford University.

Dell, Gary S., Chang, Franklin and Griffin, Zenzi M. 1993. Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science* 17, 149–195.

Dick, Judith. 1991. *A Conceptual, Case-Relation Representation of Text for Intelligent Retrieval*. Ph. D.thesis, Department of Computer Science, University of Toronto.

Dixon, Robert M. W. 1991. *A New Approach to English Grammar, on Semantic Principles*. Oxford: Clarendon Press.

Dodd, Bill J. and Zojer, Heidi. 2003. *Modern German grammar: a practical guide*. Routledge modern grammars, London: Routledge.

Dorr, Bonnie, Farwell, David, Green, Rebecca, Habash, Nizar, Helmreich, Stephen, Hovy, Eduard, Levin, Lori, Miller, Keith, Mitamura, Teruko, Rambow, Owen, Reeder, Florence and Siddharthan, Advaith. 2004. Interlingua Development and Testing through Semantic Annotation of Multilingual Text Corpora. In *COLING 2004*.

Dorr, Bonnie Jean. 1993. *Machine Translation: A View from the Lexicon*. Cambridge: MIT Press.

Dorr, Bonnie Jean. 2001. LCS Database Documentation. http://www.umiacs.umd.edu/∼ bonnie/LCS_Database_Documentation.html.

Dowty, David R. 1989. On the semantic content of the notion of Thematic Role. In Gennaro Chierchia, Barbara H. Partee and Raymond Turner (eds.), *Properties, Types and Meaning*, volume II, pages 69–130, Kluwer Academic Publishers.

Dowty, David R. 1991. Thematic Proto-roles and Argument Selection. *Language* 67(3).

Dowty, David R. 2001. The semantic asymmetry of 'argument alternations' (and why it matters). In Geart van der Meer and Alice G. B. ter Meulen (eds.), *Making Sense: From Lexeme to Discourse*, volume 44 of *Groninger Arbeiten zur germanistischen Linguistik*, Groningen: Center for Language and Cognition.

Dreyer, Hilke and Schmitt, Richard. 1985. *Lehr- und Übungsbuch der deutschen Grammatik*. Munich: Verlag für Deutsch.

Dryer, Matthew S. 1997. Are Grammatical Relations Universal? In Joan Bybee, John Haiman and Sandra Thompson (eds.), *Essays on Language Function and Language Type*, pages 115–143, Amsterdam: John Benjamins.

Dryer, Matthew S. 2005. Order of subject, object, and verb. In Martin Haspelmath, Matthew S. Dryer, David Gil and Bernard Comrie (eds.), *The World Atlas of Language Structures*, pages 330–333, Oxford: Oxford University Press.

Durrell, Martin. 2002. *Hammer's German grammar and usage*. London: Arnold.

Eco, Umberto. 1995. *The Search for the Perfect Language*. London: Fontana Press.

Edelman, Shimon and Christiansen, Morten H. 2003. How seriously should we take Minimalist syntax? A comment on Lasnik. *Trends in Cognitive Science* 7(2), 60–61.

Eysenck, Michael W. and Keane, Mark T. 1990. *Cognitive psychology: a student's handbook*. Hove: Psychology Press.

Faroqi-Shah, Yasmeen and Thompson, Cynthia K. 2003. Effect of lexical cues on the production of active and passive sentences in Broca's and Wernicke's aphasia. *Brain and Language* 85, 409–426.

Featherston, Sam. 2005. Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua* 115(11), 1525–1550.

Ferreira, Fernanda. 2000. Syntax in language production: An approach using tree-adjoining grammars. In Linda Wheeldon (ed.), *Aspects of Language Production*, Studies in cognition, Hove: Psychology Press.

Fillmore, Charles J. 1968. The Case for Case. In Bach, Emmon, Harms and Robert T (eds.), *Universals in Linguistic Theory*, pages 1–92, New York: Holt and Rinehart and Winston.

Fillmore, Charles J. 1977a. The Case for Case Reopened. In Peter Cole and Jerrold Murray Sadock (eds.), *Grammatical Relations*, pages 59–81, London: Academic Press.

Fillmore, Charles J. 1977b. Topics in Lexical Semantics. In Roger W. Cole (ed.), *Current Issues in Linguistic Theory*, London: Bloomington.

Fillmore, Charles J. 1982. Frame Semantics. In Linguistic Society of Korea (ed.), *Linguistics in the Morning Calm*, pages 111–138, Seoul: Hanshin.

Fillmore, Charles J. and Kay, Paul. 1987. The goals of Construction Grammar. Technical Report 50, University of California at Berkeley.

Fleischer, Wolfgang, Hartung, Wolfdietrich, Schildt, Joachim and Suchsland, Peter (eds.). 1983. *Kleine Enzyklopädie: Deutsche Sprache*. Leipzig: VEB Bibliographisches Institut.

Fodor, Jerry A. 1983. *Modularity of Mind*. Cambridge: MIT Press.

Frege, Gottlob. 1897. Logik. In Hans Hermes, Friedrich Kambartel and Friedrich Kaulbach (eds.), *Posthumous writings of Gottlob Frege*, volume 1979, pages pp.137–163, Oxford: Blackwell.

Fuller, Michael A. 1999. *An Introduction to Literary Chinese*. Cambridge: Harvard University Press.

Gärdenfors, Peter. 2000. *Conceptual Spaces: The Geometry of Thought*. Cambridge: MIT Press.

Garrett, Merrill F. 1984. The organization of processing structure for language production: Applications to aphasic speech. In D. Caplan, A.R. Lecours and A. Smith (eds.), *Biological perspectives on language*, pages 172–193, Cambridge: MIT Press.

Geeraerts, Dirk. 1993. Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics* 4(3), 223–272.

Genee, Inge. 1998. *Sentential complementation in a functional grammar of Irish*. Amsterdam: Holland Academic Graphics.

Goldberg, Adele. 2006. *Constructions at Work: the nature of generalization in language*. Oxford: Oxford University Press.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University Of Chicago Press.

Goldberg, Adele E., Casenhiser, Devin and Sethuraman, Nitya. 2004. Learning Argument Structure Generalizations. *Cognitive Linguistics* 14, 289–316.

Goldstein, Melvyn C., Rimpoche, Gelek and Phuntshog, Lobsang. 1991. *Essentials of Modern Literary Tibetan*. Berkeley: University of California Press.

Goodwin, William Watson. 1894. *Greek grammar*. Boston: Ginn, 1958 facsimile.

Greenbaum, Sidney. 1976. Contextual Influence on Acceptability Judgements. *Linguistics* 187, 5–11.

Greenbaum, Sidney. 1977. Judgements of Syntactic Acceptability and Frequency. *Studia Linguistica* XXXI(II).

Grimm, Laurence G. 1993. *Statistical Applications for the Behavioral Sciences*. New York: Wiley.

Grisham, R., Macleod, C. and Meyers, A. 1994. COMLEX Syntax: Building a Computational Lexicon. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages pp.268–272.

Gruber, Jeffrey S. 1965. *Studies in Lexical Relations*. Ph. D.thesis, MIT.

Gruber, Jeffrey S. 1976. *Lexical structures in syntax and semantics*, volume 25 of *North-Holland linguistic series*. Amsterdam: North-Holland Publishing Co., includes reprint of 1965 PhD thesis.

Guo Xiliang. 2003. Gu Hanyu Xuci Yanjiu Pinyi (Reflections on the Studies of Functional Words in Old Chinese). In He Da-an (ed.), *Gujin Tongsai: Hanyu de Lishi yu Fazhan (Historical Development of Chinese Language: Papers from the Third International Conference on Sinology and Linguistics Section)*, pages 49–74, Taipei: Institute of Linguistics, Academia Sinica.

Hagoort, Peter. 2003. Interplay between Syntax and Semantics during Sentence Comprehension: ERP Effects of Combining Syntactic and Semantic Violations. *Journal of Cognitive Neuroscience* 15(6), 883–899.

Hagoort, Peter, Hald, Lea, Bastiaansen, Marcel and Petersson, Karl Magnus. 2004. Integration of Word Meaning and World Knowledge in Language Comprehension. *Science* 304(438), 439–441.

Haiden, Martin. 2005. *Theta Theory*, volume 78 of *Studies in Generative Grammar*. Berlin: Mouton de Gruyter.

Halliday, M. A. K. 1985. *An Introduction to Functional Grammar*. London: Edward Arnold.

Hewings, Martin. 1999. *Advanced grammar in use: a self-study reference and practice book for advanced learners of English*. Cambridge: Cambridge University Press.

Hjelmslev, Louis. 1935. *La catégorie des cas*, volume VII of *Acta Jutlandica*. Aarhus: Aarhus University Press, 1978 Translation into Spanish by Félix Piñero Torre, Editorial Gredos, Madrid.

Hobbs, Jerry R. and Bear, John. 1990. Two Principles of Parse Preference. In Antonio Zampolli, Nicoletta Calzolari and Martha Palmer (eds.), *Current Issues in Computational Linguistics: Essays in Honour of Don Walker*, pages 503–512, Pisa, Dordrecht: Giardini Editori e Stampatori and Kluwer Academic Publishers.

Honeck, Richard P. 1973. Semantic Similarity Between Sentences. *Journal of Psycholinguistic Research* 2(2), 137–151.

Ioannidis, John P. A. 2005a. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 294(2), 218–228.

Ioannidis, John P. A. 2005b. Why Most Published Research Findings Are False. *PLoS Medicine* 2(8).

Jackendoff, Ray S. 1972. *Semantic Interpretation in Generative Grammar*. Cambridge: MIT Press.

Jackendoff, Ray S. 1983. *Semantics and Cognition*. Cambridge: MIT Press.

Jackendoff, Ray S. 1990. *Semantic Structures*. Cambridge: MIT Press.

Jackendoff, Ray S. 1992. *Languages of the Mind: Essays on Mental Representation*. Cambridge: MIT Press.

Jamieson, Susan. 2004. Likert scales: how to (ab)use them. *Medical Education* 38(12), 1217–1218.

Jespersen, Otto. 1924. *The philosophy of grammar*. London: Allen & Unwin.

Kahrs, Eivind Georg. 1999. *Indian Semantic Analysis: The Nirvacana Tradition*. Cambridge: Cambridge University Press.

Kail, Michele. 2004. On-line grammaticality judgments in French children. *Journal of Child Language* 31, 713–737.

Kaplan, Robert. M. and Bresnan, Joan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In J. Bresnan (ed.), *The Mental Representation of Grammatical Relations*, pages 173–281, Cambridge, MA: MIT Press.

Kaplan, Robert M., Bush, James W. and Berry, Charles C. 1979. Health Status Index: Category Rating versus Magnitude Estimation for Measuring Levels of Well-Being. *Medical Care* 17(5), 501–525.

Karabanov, Anke, Bosch, Peter and König, Peter. 2007. Eye Tracking as a tool for investigating the comprehension of referential expressions. In Sam Featherston and Wolfgang Sternefeld (eds.), *Roots: Linguistics in Search of its Evidential Base*, Amsterdam: Mouton de Gruyter.

Karlgren, Bernhard. 1920. Le proto-chinois, langue flexionelle. *Journal Asiatique* 15, 205–232.

Katz, Jerrold J. and Fodor, Jerry A. 1963. The structure of a semantic theory. *Language* 2(39), 170–210.

Katz, Jerrold J. and Postal, Paul M. 1964. *An Integrated Theory of Linguistic Descriptions*. Cambridge: MIT Press.

Keenan, Edward L. 1976. Towards a universal definition of subject. In Charles N. Li (ed.), *Subject and topic*, pages 303–333, New York: Academic Press.

Keenan, Edward L. 1985. Passive in the World's Languages. In Timothy Shopen (ed.), *Language typology and syntactic description*, pages 243–281, Cambridge: Cambridge University Press.

Keller, Frank. 2003. A Psychophysical Law for Linguistic Judgments. In Richard Alterman and David Kirsh (eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 652–657, Boston.

Keller, Frank. 2006. Linear Optimality Theory as a Model of Gradience in Grammar. In Gisbert Fanselow, Caroline Féry, Ralph Vogel and Matthias Schlesewsky (eds.), *Gradience in Grammar: Generative Perspectives*, Oxford University Press.

Keller, Frank and Alexopoulou, Theodora. 2001. Phonology Competes with Syntax: Experimental Evidence for the Interaction of Word Order and Accent Placement in the Realization of Information Structure. *Cognition* 79(3), 301–372.

Keller, Frank and Asudeh, Ash. 2001. Constraints on Linguistic Coreference: Structural vs. Pragmatic Factors. In Johanna D. Moore and Keith Stenning (eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 483–488, Mahawah: Lawrence Erlbaum.

Kilgarriff, Adam. 1997a. The hard parts of lexicography. *International Journal of Lexicography* 11(1), 51–54.

Kilgarriff, Adam. 1997b. I Don't Believe in Word Senses. *Computers and the Humanities* 31(2), 91–113.

Kilgarriff, Adam. 2007. Googleology is Bad Science. *Computational Linguistics* 33(1), 147–151.

Kim, Albert and Osterhout, Lee. 2005. The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language* 52, 205–225.

Kita, Sotaro. 2000. How Representational Gestures Help Speaking. In David MacNeill (ed.), *Language and Gesture*, pages 162–85, Cambridge University Press.

Krifka, Manfred. 2004. Semantic and pragmatic conditions for the Dative Alternation. *Korean Journal of English Language and Linguistics* 4, 1–32.

Kruijff-Korbayová, Ivana and Steedman, Mark. 2003. Discourse and Information Structure. *Journal of Logic, Language and Information* 12, 249–259.

Kuno, Susumo. 1976. Subject, theme, and the speaker's empathy – a reexamination of relativization phenomena. In Charles Li (ed.), *Subject and Topic*, pages 417–44, New York: Academic Press.

Kuperberg, G. R., McGuire, P. K., Bullmore, E. T., Brammer, M. J., Rabehesketh, S., Wright, I. C., Lythgoe, D. J., Williams, S. C. R. and David, A. S. 2000. Common and Distinct Neural Substrates for Pragmatic, Semantic, and Syntactic Processing of Spoken Sentences: An fMRI Study. *Journal of Cognitive Neuroscience* 12(2), 321–341.

La Heij, Wido. 1988. Components of Stroop-like interference in picture naming. *Memory and Cognition* 16, 400–410.

Laka, Itziar. 2008. *A brief grammar of Euskara, the Basque language*. Bilbao: Argitarapen Zerbitzua.

Lakoff, George. 1971. On Generative Semantics. In D. D. Steinberg and L. A. Jakobovits (eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology*, pages 232–296, Cambridge: Cambridge University Press.

Landauer, Thomas K, Foltz, Peter W. and Laham, Darrell. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes* 25, 259–284.

Langacker, Ronald W. 1990. *Concept, Image and Symbol: the Cognitive Basis of Grammar*. Berlin: Mouton de Gruyter.

LaPolla, Randy J. 1993. Arguments Against 'Subject' and 'Direct Object' as Viable Concepts in Chinese. *Bulletin of the Institute of History and Philology, Academia Sinica* .

Larson, Mildred L. 1984. *Meaning-based translation: a guide to cross-language equivalence*. Lanham: University Press of America.

Lawless, Harry T. 1989. Logarithmic Transformation of Magnitude Estimation Data and Comparisons of Scaling Methods. *Journal of Sensory Studies* 4(2), 75–86.

Lee, Thomas Hun-Tak. 1996. Theoretical Issues in Language Development and Chinese Child Language. In Cheng teh James Huang and Yen hui Audrey Li (eds.), *New horizons in Chinese linguistics*, Dordrecht: Kluwer.

Lee Joon Ho, Kim Myoung Ho and Lee Yoon Joon. 1993. Information Retrieval Based on Conceptual Distance in IS-A Hierarchies. *Journal of Documentation* 49(2), 188–207.

Leech, Geoffrey, Garside, Roger and Bryant, Michael. 1994. Claws4: The Tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, pages 622–628, Kyoto.

Legendre, Géraldine, Miyata, Yoshiro and Smolensky, Paul. 1990. Harmonic Grammar – A formal multi-level connectionist theory of linguistic well-formedness. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 884–891.

Legendre, Géraldine, Sorace, Antonella and Smolensky, Paul. 2006. The Optimality Theory-Harmonic Grammar connection. In Paul Smolensky and Géraldine Legendre (eds.), *The Harmonic Mind*, Chapter 20, pages 903–966, MIT Press.

Levelt, Willem J. M. 2001. Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences of the United States of America* 98(23), 13464–13471.

Levin, Beth. 1993. *English Verb Classes and Alternations*. Chicago: University of Chicago Press.

Levin, Beth and Rappaport Hovav, Malka. 1996. From Lexical Semantics to Argument Realization, unpublished manuscript.

Levin, Beth and Rappaport Hovav, Malka. 2005. *Argument Realization*. Research Surveys in Linguistics, Cambridge: Cambridge University Press.

Levin, Magnus. 2006. Collective nouns and language change. *English Language and Linguistics* 10(2), 321–343.

Li, Charles N. and Thompson, Sandra A. 1997. *Mandarin Chinese: A functional reference grammar*. Taipei: Crane Publishing.

Li Dejin and Cheng Meizhen. 1988. *Waiguoren shiyong hanyu yufa (A practical Chinese grammar for foreigners)*. Beijing: Sinolingua.

Liljencrants, Johan and Lindblom, Bjorn. 1972. Numerical Simulation of Vowel Quality Systems: The Role of Perceptual Contrast. *Language* 48(4), 839–862.

Lily, William and Colet, John. 1549. *A short introduction of grammar*, volume 262 of *English linguistics, 1500-1800, a collection of facsimile reprints*. Menston: Scholar Press, 1970 facsimile reprint of 1549 edition.

Lin Dekang. 1998. A dependency-based method for evaluating broad-coverage parsers. *Natural Language Engineering* 4, 97–114.

Liu Feng-hsi. 1997. An Aspectual Analysis of ba. *Journal of East Asian Linguistics* 6(1), 51–99.

Lodge, Milton. 1981. *Magnitude scaling: quantitative measurement of opinions*. Beverly Hills: Sage.

Love, Tracy and Oster, Elizabeth. 2002. On the Categorization of Aphasic Typologies: The SOAP (A Test of Syntactic Complexity). *Journal of Psycholinguistic Research* 31(5).

Lü Yiwei. 1311. *Yuzhu*. Beijing: Zhonghua Press, reprinted 1988 as Zhuyuci Jizhu (Variorum).

Lyons, John. 1968. *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.

MacNeill, David. 2005. *Gesture & Thought*. University of Chicago Press.

Mahon, Bradford Z., Costa, Albert, Peterson, Robin and Vargas, Kimberly A. 2007. Lexical selection is not by competition: a reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of experimental psychology. Learning, memory, and cognition* 33(3), 503–535.

Mann, William C. and Thompson, Sandra A. 1988. Rhetorical Structure Theory: Towards a functional theory of text organization. *Text* 8(3), 243–281.

Manning, Chris and Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Marcus, Mitchell, Kim, Grace, Marcinkiewicz, Mary Ann, MacIntyre, Robert, Bies, Ann, Ferguson, Mark, Katz, Karen and Schasberger, Britta. 1994. The Penn TreeBank: Annotating predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*.

Marx, Edeltrud. 1999. Gender Processing in Speech Production: Evidence from German Speech Errors. *Journal of Psycholinguistic Research* 28(6), 601–621.

McClelland, James L. and Rogers, Timothy T. 2003. The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience* 4, 310–322.

McDonald, Janet L. 2000. Grammaticality judgments in a second language: Influences of age. *Applied Psycholinguistics* 21, 395–423.

McEnery, Tony and Xiao, Richard. 2005. Passive constructions in English and Chinese: a corpus-based contrastive study. In *Online Proceedings of Corpus Linguistics 2005*, Birmingham University.

Michell, Joel. 1986. Measurement scales and statistics: a clash of paradigms. *Psychological Bulletin* 100(3), 398–407.

Mihalcea, Rada and Csomai, Andras. 2005. SenseLearner: Word Sense Disambiguation for All Words in Unrestricted Text. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor.

Miller, George A., Beckwith, Richard, Fellbaum, Christiane, Gross, Derek and Miller, Katherine. 1990. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography* 3(4), 235–244.

Miller, George A. and Charles, Walter G. 1991. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes* 6(1), 1–28.

Miyao, Yusuke and Tsujii, Jun'ichi. 2005. Probabilistic disambiguation models for wide-coverage HPSG parsing. In *Proceedings of of 43rd ACL*, pages 83–90.

Moens, Marc and Steedman, Mark. 1988. Temporal ontology and temporal reference. *Computational Linguistics* 14(2), 15–28.

Moore, David S. and McCabe, George P. 2003. *Introduction to the Practice of Statistics*. New York: Freeman & Co.

Murphy, Brian and Vogel, Carl. 2007. Statistically-Constrained Shallow Text Marking: Techniques, Evaluation Paradigm and Results. In Edward J. Delp III and Ping Wah Wong (eds.), *Proceedings of Security, Steganography, and Watermarking of Multimedia Contents IX*, San José: SPIE.

Murphy, Brian, Vogel, Carl and Opitz, Conny. 2005. Cross-linguistic Empirical Analysis of Constraints on Passive, presentation to the Symposium on Interdisciplinary Themes in Cognitive Language Research.

Murphy, Victoria A. 1997. The effect of modality on a grammaticality judgement task. *Second Language Research* 13(1), 34–65.

Myers, James. 2007. MiniJudge: Software for small-scale experimental syntax. *International Journal of Computational Linguistics and Chinese Language Processing* 12(2), 175–194.

Nakamura, Hajime. 1960. *Ways of thinking of Eastern peoples: India, China, Tibet and Japan*. London: Kegan Paul, revised 1997 translation from the japanese edition.

Newman, Aaron J., Pancheva, Roumyana and Ozawa, Kaori. 2001. An Event-Related fMRI Study of Syntactic and Semantic Violations. *Journal of Psycholinguistic Research* 30(3), 339–364.

Ni, W., Constable, R. T., Mencl, W. E., Pugh, K. R., Fulbright, R. K., Shaywitz, S. E., Shaywitz, B. A., Gore, J. C. and Shankweiler, D. 2000. An Event-related Neuroimaging Study Distinguishing Form and Content in Sentence Processing. *J. Cognitive Neuroscience* 12(1), 120–133.

Nicol, Janet L. and Love, Tracey. 2000. Overarching agrammatism: When comprehension involves production. In Y. Grodzinsky and L. Shapiro (eds.), *Language and the Brain: Representation and Processing*, Chapter 5, pages 105–120, New York: Academic Press.

Nieuwland, Mante S. and Berkum, Jos J. A. Van. 2006. When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. *Journal of Cognitive Neuroscience* 18, 1098–1111.

Nieuwland, Mante S., Otten, Marte and Berkum, Jos J. A. Van. 2007. Who are You Talking About? Tracking Discourse-level Referential Processing with Event-related Brain Potentials. *Journal of Cognitive Neuroscience* 19, 228–236.

Noonan, Michael. 2004. Subjectless clauses in Irish. In Peri Bhaskararao and Karumuri Venkata Subbarao (eds.), *Non-nominative Subjects*, New York: John Benjamins.

Nyberg, Eric and Mitamura, Teruko. 2000. The KANTOO Machine Translation Environment. In *Proceedings of AMTA 2000*, Mexico City.

Onitsuka, Toshiaki, Shenton, Martha E., Kasai, Kiyoto, Nestor, Paul G., Toner, Sarah K., Kikinis, Ron and Jolesz, Ferenc A. 2003. Fusiform Gyrus Volume Reduction and Facial Recognition in Chronic Schizophrenia. *Archives of General Psychiatry* 60, 349–355.

Orth, Bernhard and Wegener, Bernd. 2006. Scaling occupational prestige by magnitude estimation and category rating methods: A comparison with the sensory domain. *European Journal of Social Psychology* 13(4), 417–431.

Ortiz de Urbina, Jon. 2002. Focus of Correction and Remnant Movement in Basque. In Xabier Artiagoitia Beaskoetxea, Patxi Goenaga Mendizabal and Joseba A. Lakarra Andrinua (eds.), *Erramu Boneta: Festscrift for Rudolf P.G. de Rijk*, pages 510–524, Bilbao: University of the Basque Country.

Osterhout, Lee and Nicol, Janet. 1999. On the distinctiveness, independence, and time course of the brain responses to syntactic and semantic anomalies. *Language and Cognitive Processes* 14, 283–317.

Ostler, Nicholas. 1980. *A theory of case linking and agreement*. Bloomington: Indiana University Linguistics Club, truncated version of unpublished 1979 MIT PhD.

Ouhalla, Jamal. 1999. *Introducing transformational grammar: from principles and parameters to minimalism*. London: Arnold.

Owens, Jonathan. 1984. Structure, Class and Dependency: Modern linguistic theory and the Arabic grammatical tradition. *Lingua* 64, 25–62.

Özyürek, Asli, Willems, Roel M., Kita, Sotaro and Hagoort, Peter. 2007. On-line Integration of Semantic Information from Speech and Gesture: Insights from Event-related Brain Potentials. *Journal of Cognitive Neuroscience* 19(4), 605–616.

Paik, T. S. 1974. Review of A Reader on the Sanskrit Grammarians by J. F. Staal. *Language* 50(3), 591–598.

Palmer, Frank Robert. 1994. *Grammatical roles and relations*. Cambridge textbooks in linguistics, Cambridge: Cambridge University Press.

Palmer, Martha. 2000. Consistent Criteria for Sense Distinctions. *Computers and the Humanities (special issue on SENSEVAL98: Evaluating Word Sense Disambiguation Systems)* 34(1-2).

Partee, Barbara H. 1965. *Subject and Object in English*. Ph. D.thesis, MIT.

Partee, Barbara H. 1979. Semantics: mathematics of psychology? In Rainer Bauerle, Urs Egli and Armin von Stechow (eds.), *Semantics from Different Points of View*, Berlin: Springer Verlag.

Peterson, Robert R. and Savoy, Pamela. 1998. Lexical Selection and Phonological Encoding During Language Production: Evidence for Cascaded Processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24(3), 539–557.

Peyraube, Alain. 2001. Some reflections on the sources of the *Mashi Wentong*. In Michael Lackner, Iwo Amelung and Joachim Kurtz (eds.), *New Terms for New Ideas: Western Knowledge and Lexical Change in Late Imperial China*, pages 341–356, Leiden: Brill.

Piñango, Maria Mercedes. 2006. Thematic roles as event structure relations. In Ina Bornkessel, Matthias Schlesewsky, Bernard Comrie and Angela D. Friederici (eds.), *Semantic Role Universals and Argument Linking: Theoretical, Typological and Psycholinguistic Perspectives*, pages 303–326, Mouton de Gruyter.

Pinker, Stephen. 1989. *Learnability and cognition. The acquisition of argument structure*. Cambridge: MIT Press.

Piwek, Paul and van Deemter, Kees. 2006. Constraint-based Natural Language Generation: A Survey. Technical Report, Computing Department and The Open University.

Pollard, Carl. 1994. Toward a Unified Account of Passive in German. In John Nerbonne, Klaus Netter and Carl Pollard (eds.), *German in Head-Driven Phrase Structure Grammar*, pages 273–296, CSLI Publications.

Pollard, Carl and Sag, Ivan A. 1994. *Head-Driven Phrase Structure Grammar*. Stanford: CSLI.

Porter, David W. (ed.). 2002. *Excerptiones de Prisciano: the source for Aelfric's Latin-Old English grammar*. New York: D.S. Brewer.

Primus, Beatrice. 1999. *Case and Thematic Roles: Ergative, Accusative and Active*. Tübingen: Max Niewmeyer Verlag.

Primus, Beatrice. 2002. Proto-roles and case-selection in optimality theory. Technical Report, Universität Köln.

Primus, Beatrice. 2006. Hierarchy mismatches and the dimensions of role semantics. In Ina Bornkessel, Matthias Schlesewsky, Bernard Comrie and Angela D. Friederici (eds.), *Semantic Role Universals and Argument Linking: Theoretical, Typological and Psycholinguistic Perspectives*, pages 53–88, Mouton de Gruyter.

Prince, Alan and Smolensky, Paul. 1997. Optimality: From Neural Networks to Universal Grammar. *Science* 275(5306), 1604–1610.

Prince, Ellen P. 1981. Toward a taxonomy of given-new information. In Peter Cole (ed.), *Radical Pragmatics*, pages 223–256, New York: Academic Press.

Pulleyblank, Edwin George. 1960. Studies in Early Chinese Grammar, Part I. *Asia Major* 8(1), 36–67.

Quine, Willard V. 1960. *Word and object*. Cambridge: M.I.T. Press.

Quine, Willard V. 1964. Meaning and Translation. In Jerrold J. Katz and Jerry A. Fodor (eds.), *The Structure of Language: readings in the philosophy of language*, pages 460–478, New York: Prentice Hall.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey and Svartnik, Jan. 1985. *A Comprehensive grammar of the English language*. London: Longman.

Rappaport, Malka and Levin, Beth. 1988. What to do with θ-Roles? In Wendy Wilkins (ed.), *Thematic Relations*, pages 7–36, New York: Academic Press.

Rayner, Keith. 1998. Eye movements in reading and information processing: Twenty years of research. *Psychological Bulletin* 124(3), 372–422.

Regier, Terry, Kay, Paul and Khetarpal, Naveen. 2007. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences of the United States of America* 104, 1436–1441.

Reinhart, Tanya. 2002. The Theta System – an overview. *Theoretical Linguistics* 28(3).

Robins, Robert H. 1987. Polysemy and the Lexicographer. In Robert Burchfield (ed.), *Studies in Lexicography*, Oxford: Oxford University Press.

Robinson, Ian. 1975. *The new grammarians' funeral: a critique of Noam Chomsky's linguistics*. Cambridge: Cambridge University Press.

Rocher, Rosane. 1968. *La théorie des voix du verbe dans l'école paninéenne*. Brussels: Presses universitaires de Bruxelles.

Rosch, Eleanor. 1973. Natural Categories. *Cognitive Psychology* 7, 328–350.

Russell, Bertrand. 1946. *History of Western Philosophy*. London: Routledge.

Russell, Melanie, Corley, Martin and Lickley, Robin J. 2005. Magnitude estimation of disfluency by stutterers and nonstutterers. *Phonological Encoding and Monitoring in Normal and Pathological Speech* 1(5), 248–260.

Saeed, John I. 1997. *Semantics*. Oxford: Blackwell.

Sampson, Geoffrey. 1995. *English for the computer: the Susanne corpus and analytic scheme*. Oxford: Clarendon Press.

Sanfilippo, Antonio. 1990. *Grammatical Relations, Thematic Roles and Verb Semantics*. Ph. D.thesis, Centre for Cognitive Science, University of Edinburgh.

Sapir, Edward. 1921. *Language: an introduction to the study of speech*. New York: Harcourt and Brace.

Sapir, Edward. 1931. The Function of an International Auxiliary Language. In David G. Mandelbaum (ed.), *Selected writings of Edward Sapir in language, culture, and personality*, Berkeley: University of California Press.

Sapir, Edward, Swadesh, Morris and Morris, Alice V. 1932. *The Expression of the Ending-Point Relation in English, French and German*. Language Monographs, Baltimore: Waverly Press.

Schank, Robert and Abelson, R. 1977. *Scripts, Plans, Goals and Understanding*. Hillsdale: Erlbaum.

Schlegel, Friedrich. 1808. *Ueber die Sprache und Weisheit der Indier*. Heidelberg: Mohr und Zimmer, 1995 Routledge facsimile.

Schütze, Carson T. 1996. *The Empirical Basis of Linguistics*. Chicago: University of Chicago Press.

Schütze, Carson T. 2005. Thinking about what we are asking speakers to do. In Stephan Kepser and Marga Reis (eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives*, pages 457–485, Mouton de Gruyter.

Schütze, Hinrich. 1993. Word space. In Stephen José Hanson, Jack D. Cowan and C. Lee Giles (eds.), *Advances in Neural Information Processing Systems*, volume 6, pages 895–902, San Mateo: Morgan Kaufmann.

Shapiro, Mary. 1997. *Style Shifting: The perception and production of formality in English*. Ph. D.thesis, Department of Linguistics, University of Texas at Austin.

Siegel, Sidney and Castellan, N. John. 1988. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Slobin, Dan I. and Bever, Thomas G. 1982. Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition* 12, 229–265.

Smith, Carlotta. 1991. *The Parameter of Aspect*. Dordrecht: Kluwer.

Smith, Michael. 1987. *The Semantics of Dative and Accusative in German: An investigation in cognitive grammar*. Ph. D.thesis, University of California, San Diego, Linguistics Department.

Smith, Neil. 1999. *Chomsky: ideas and ideals*. Cambridge: Cambridge University Press.

Somers, Harold. 1987. *Valency and Case in Computational Linguistics*. Edinburgh: Edinburgh University Press.

Sorace, Antonella. 1992. *Lexical conditions on Syntactic Knowledge: Auxiliary Selection in Native and Non-Native Grammars of Italian*. Ph. D.thesis, University of Edinburgh, Edinburgh.

Sorace, Antonella and Keller, Frank. 2005. Gradience in Linguistic Data. *Lingua* 115(11), 1497–1524.

Southwood, M. Helen and Flege, James E. 1999. Scaling foreign accent: direct magnitude estimation versus interval scaling. *Clinical Linguistics and Phonetics* 13(5), 335–349.

Sowa, John F. 2000. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. London: Brooks/Cole.

Spencer, N. J. 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* 2(2), 83–98.

Staal, Johan Frederik. 1972. *A reader on the Sanskrit grammarians*, volume 1 of *Studies in linguistics*. Cambridge: MIT Press.

Steedman, Mark. 2007. Information-structural semantics for English intonation. In Lee Chungmin, Matthew Gordon and Daniel Büring (eds.), *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation (proceedings of the LSA 2001 Summer Institute Workshop on Topic and Focus)*, pages 245–264, Dordrecht: Springer.

Sterne, Jonathan A. C. and Smith, George Davey. 2001. Sifting the evidence – what's wrong with significance tests? *British Medical Journal* 322, 226–231.

Stevens, Stanley Smith. 1946. On the Theory of Scales of Measurement. *Science* 103, 677–680.

Stevens, Stanley Smith. 1975. *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. New York: John Wiley & Sons.

Swadesh, Morris. 1972. *The origin and diversification of language*. London: Routledge.

Tabor, Whitney. 1993. The gradualness of syntactic change: a corpus proximity model. In *XIth International Conference on Historical Linguistics*, Los Angeles: University of California.

Tai, James H-Y. and Chan, Marjorie K. M. 1999. Some reflections on the periodization of the Chinese language. In Alain Peyraube and Chaofen Sun (eds.), *Studies in Chinese Historical Syntax and Morphology: Linguistic Essays in Honor of Mei Tsu-lin*, pages 223–239, Paris: École des Hautes Études en Sciences Sociales.

Talmy, Leonard. 1985. Lexicalization Patterns: Semantic Structures in Lexical Form. In Timothy Shopen (ed.), *Language Typology And Symbolic Description*, volume 3, Cambridge: Cambridge University Press.

Tanenhaus, Michael K., Burgess, Curt, Hudson-D'Zmura, Susan B. and Carlson, Greg N. 1987. Thematic roles in language processing. In *Proceedings of the Ninth Annual Cognitive Science Society*, pages 587–596, Hillsdale: LEA Press.

Tanenhaus, Michael K., Spivey-Knowlton, Michael J., Eberhard, Kathleen M. and Sedivy, Julie C. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268(5217), 1632–1634.

Ting Jen. 1998. Deriving the Bei-construction in Mandarin Chinese. *Journal of East Asian Linguistics* 7(4), 319–354.

Traxler, Matthew J. and Pickering, Martin J. 1996. Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language* 35, 454–475.

Tsao Feng-fu. 1987. A Topic-Comment Approach to the Ba Construction. *Journal of Chinese Linguistics* 15, 1–54.

Uchida, Hiroshi and Meiying, Zhu. 2001. The Universal Networking Language: Beyond Machine Translation. In *Proceedings of International Symposium on Language in Cyberspace 2001*, Seoul.

UNL Centre. 2001. The Universal Networking Language Specifications 3.0. http://www.unl.ias.unu.edu/unlsys/unl/UNL Specifications.htm.

Uylings, H. B. M., Malofeeva, L. I., Bogolepova, I. N., Jacobsen, A. M., Amunts, K. and Zilles, K. 2005. No postnatal doubling of number of neurons in human Broca's areas (Brodmann areas 44 and 45)? A stereological study. *Neuroscience* 136, 715–728.

Vallduvi, Enric and Engdahl, Elisabet. 1996. The Linguistic Realisation of Information Packaging. *Linguistics* 34, 459–519.

van Valin, Robert D. 1999. Generalized Semantic Roles and the Syntax-Semantics Interface. In F. Corblin, C. Dobrovie-Sorin and J. M. Marandin (eds.), *Empirical Issues in Formal Syntax and Semantics*, pages 373–389, The Hague: Thesus.

van Valin, Robert D. unpublished. A Summary of Role and Reference Grammar. http://wings.buffalo.edu/linguistics/rrg/rrg/RRGsummary.pdf.

van Valin, Robert D. and LaPolla, Randy J. 1997. *Syntax: Structure, Meaning and Function*. Cambridge: Cambridge University Press.

Velleman, Paul F. and Wilkinson, Leland. 1993. Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician* 47(1), 65–72.

Vendler, Zeno. 1967. Verbs and times. *Linguistics in Philosophy* pages 97–121.

Verhagen, Pieter C. 2001. *A history of Sanskrit grammatical literature in Tibet*, volume 2: Assimilation into indigenous scholarship. Leiden: Brill.

Verkuyl, Henk J. 1993. *A Theory of Aspectuality*. Cambridge Studies in Linguistics, Cambridge University Press.

Vickers, Zata M. 1983. Magnitude Estimation vs Category Scaling of the Hedonic Quality of Food Sounds. *Journal of Food Science* 48(4), 1183–1186.

Vigliocco, Gabriella and Hartsuiker, Robert J. 2002. The Interplay of Meaning, Sound, and Syntax in Sentence Production. *Psychological Bulletin* 128(3), 442–472.

Vogel, Carl and Cooper, Robin. 1995. Robust Chart Parsing with Mildly Inconsistent Feature Structures. In Andreas Schoter and Carl Vogel (eds.), *Natural Language Understanding and Logic Programming*, volume 10 of *Working Papers in Cognitive Science*, pages 197–216, University of Edinburgh.

Šarčevic, Susan. 1997. *New approach to legal translation*. The Hague: Kluwer Law International.

Vygotsky, Lev Semenovich. 1934. *Thought and Language*. Cambridge: MIT Press, 1962 translation into English.

Wang Hai. 1992. *Chinese-English Dictionary of Function Words*. Beijing: Sinolingua.

Wasow, Thomas and Arnold, Jennifer. 2005. Intuitions in linguistic argumentation. *Lingua* 115, 1481–1496.

Wechsler, Stephen. 1991. *Argument Structure and Linking*. Ph. D.thesis, Stanford University.

Wegener, Bernd. 1983. Category-Rating and Magnitude Estimation Scaling Techniques: An Empirical Comparison. *Sociological Methods & Research* 12(1), 31–75.

Weiner, E. Judith and Labov, William. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19, 29–58.

Welch, Bernard Lewis. 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika* 29, 350–361.

Whissell, Cynthia. 1989. The Dictionary of Affect in Language. In Robert Plutchik and Henry Kellerman (eds.), *Emotion: theory, research and experience*, volume 4, The measurement of emotions, London: Academic Press.

Whitney, William Dwight. 1889. *Sanskrit Grammar*. Cambridge: Harvard University Press, 1987 facsimile.

Wierzbicka, Anna. 1996. *Semantics: Primes and universals*. Oxford: Oxford University Press.

Wilkins, John. 1668. *An essay towards a real character, and a philosophical language*. English linguistics, 1500-1800; a collection of facsimile reprints; no. 119, 1968, Scholar Press.

Wilks, Yorick. 1975. An intelligent analyzer and understander of English. *Communications of ACM* 18(5), 264–274.

Wolf, Ernest M. 1960. The German Equivalents of English ‘to like’. *The German Quarterly* pages 49–53.

Wulfeck, Beverly B. 1993. A Reaction Time Study of Grammaticality Judgments in Children. *Journal of Speech and Hearing Research* 36, 1208–1215.

Wunderlich, Dieter. 2006. Argument hierarchy and other factors determining argument realization. In Ina Bornkessel, Matthias Schlesewsky, Bernard Comrie and Angela D. Friederici (eds.), *Semantic Role Universals and Argument Linking: Theoretical, Typological and Psycholinguistic Perspectives*, Mouton de Gruyter.

Wygant, Steve and Lindorf, Ron. 1999. Surveying collegiate Net surfers. *Quirk's Marketing Research Review* .

Yip Po-Ching and Rimmington, Don. 1997. *Chinese: An Essential Grammar*. London: Routledge.

Zagona, Karen. 2002. *The syntax of Spanish*. Cambridge: Cambridge University Press.

Zifonun, Gisela, Hoffmann, Ludger, Strecker, Bruno and Ballweg, Joachim. 1997. *Grammatik der deutschen Sprache*, volume 7 of *Schriften des Instituts für Deutsche Sprache*. Berlin: W. de Gruyter.