Statistically-constrained shallow text marking: techniques, evaluation paradigm and results

Brian Murphy & Carl Vogel*
Computational Linguistics Lab, Trinity College Dublin, Dublin 2, Ireland

ABSTRACT

We present three natural language marking strategies based on fast and reliable shallow parsing techniques, and on widely available lexical resources: lexical substitution, adjective conjunction swaps, and relativiser switching. We test these techniques on a random sample of the British National Corpus. Individual candidate marks are checked for goodness of structural and semantic fit, using both lexical resources, and the web as a corpus. A representative sample of marks is given to 25 human judges to evaluate for acceptability and preservation of meaning. This establishes a correlation between corpus based felicity measures and perceived quality, and makes qualified predictions. Grammatical acceptability correlates with our automatic measure strongly (Pearson's $r = 0.795$, $p = 0.001$), allowing us to account for about two thirds of variability in human judgements. A moderate but statistically insignificant (Pearson's $r = 0.422$, $p = 0.356$) correlation is found with judgements of meaning preservation, indicating that the contextual window of five content words used for our automatic measure may need to be extended.

Keywords: information hiding, shallow parsing, web corpus, human judgement, correlation

1. INTRODUCTION

Within the problem of information hiding in texts, we address the issue of evaluating techniques for making meaning-preserving transforms to plain text that will pass human and machine detection attacks. The methods we suggest are not generally reversible, preventing reliable wholesale mark removal, and thus making them applicable to watermarking applications such as intellectual ownership assertion and traitor tracing. Past research has addressed deep structure techniques and methods based on lexical relations (predominantly, synonymy) [1] and movement of phrasal constituents. [2–4] We suggest here an evaluation method, and apply that method to an alternative hypothesis of shallow transformational processing (in a sense to be detailed in the following section).

Three desiderata that we apply to candidate transform methods is that they should be: undetectable, preserving of sentence acceptability and meaning preserving. [5] These are independent criteria. A transform of a sentence can preserve grammaticality without preserving meaning, and vice versa. Meaning and grammaticality can be preserved while statistical likelihood dramatically altered, for example, rendering discovery possible with *n*-gram based-attacks. [6] Our transformation method appeals to corpus driven techniques in establishing an on-line assessment of applicability of a specific transform, and should thus be indetectable on the basis of frequency oriented automated statistical processing.

2. SHALLOW PARSE MARKING TECHNIQUES

Some extant work operates on the assumption that reliable parsing is available for the foundation of deep structural analysis involved in adverbial movement and alternations such passive-active. However, constituent boundary detection is a significant problem even when all that is at stake is determining noun-phrase content, given the possibility of prepositional phrase, relative clauses, appositive, etc. within a nominal constituent. For example a state of the art deep syntactic parser [7] can achieve about 90% accuracy per node of a syntactic or dependency structure. Since a construction such as the passive, which has been suggested as a possible linguistic structure for watermarking, [2, 3, 8] must be represented with at lease five head nodes (subject, auxiliary verb, main verb, *by*-phrase, embedded object), we can expect approximately 60% success in detecting it (0.9 to the power of five).

Murphy and Vogel[5] attempted to evaluate the viability of structural transforms assumed in constituent-movement methods and established the reliability of specific classes of transforms, by construction type, under the assumption of a 'perfect' parser and generator. That is, setting aside whether reliably automated processing can detect candidates for marking,

*vogel@cs.tcd.ie; brian.murphy@unitn.it

a method of evaluating which transforms would be safe in the sense of preserving meaning and well-formedness was provided. However, automated deep analysis of the kind needed (also by approaches such as Atallah et al.[4]) does not appear sufficiently robust.

Here, we propose three shallow-processing techniques which are thus insensitive to the brittleness of deep processing and which benefit from (qualified) sensitivity of human judges to frequency effects. All three techniques rely on part-of-speech (POS) tagging. We use the Brill tagger, which has achieved 97% accuracy[9,10] in an evaluation on balanced British English text; thus, we do not eschew syntactic analysis in its entirety, but remain very close to the level of lexical items in closed and open class word categories (we do examine multi-word expressions).

The first method is representative of function-word near-synonymy relations. We search for the pattern "COMMON-NOUN *who*" or "COMMON-NOUN *which*", and replace the relativiser with *that* as seen in examples (1) and (2).[†]

(1) The only depôt ~~which~~ that had sufficient clearance ...

(2) ... a way of competing with women ~~who~~ that are brighter ...

This is not reliably reversible, because it is impossible to determine which of the instances of *that* in a marked text are original, and which substituted (unless seeded by a fixed key). Also it is not always possible to decide whether *who* or *which* is the correct alternate of *that*, since it may not be possible to determine the animacy/humanity of the relativiser head noun (*depôt* and *women* in these examples), particularly for words that represent organisations of people. In English these may use *which*, or the personal form, *who*.

The pattern "ADJECTIVE CONJUNCTION ADJECTIVE COMMON-NOUN" is the pivot for the second method, swapping adjective positions, as in (3).

(3) ... the "chasm" between ~~lawful and unlawful~~ unlawful and lawful acts ...

As with the first method, if the swap is feasible (see §5 for assessment of these methods), it should not be possible for an attacker to detect swapped adjective conjunction constructions, as opposed to originals. However, it would be possible to scramble these marks for mark destruction.

The third method we report on here considers individual content words (adjectives (4), verbs (5), nouns (6, 7) and adverbs (8)) to identify likely lexical substitutions using WordNet.[11] This differs from previous approaches[1,12] which use hand-crafted lexical substitution sets based on existing synonym sets. We use a wider range of lexical relations (synonymy, hyponymy and hypernymy) from WordNet in a fully automatic fashion, using corpus frequency data and graph topography to select only lexical substitutes that are highly likely to be replacements of the original text word in question. Word sense disambiguation is unnecessary, as our system suggests only related words that are a likely replacement for most or all of the original word's senses.

(4) Excellent quality wines of ~~superb~~ brilliant elegance and great finesse.

(5) The Rayner Review ~~stated~~ said: Particular attention needs to be paid ...

(6) ... and escape into the ~~vastness~~ enormity of the tundra ...

(7) They strip the leaves and often the bark off all the trees and ~~shrubs~~ bushes they can reach ...

(8) The dunderheads etc. wander in and out picking ~~ravenously~~ hungrily at the remains of a beetroot salad.

[†] The original text is shown italicized, and replacement text underlined. Unless otherwise noted, all examples are from the British National Corpus.

In this method, a word to be substituted provides links to alternative SynSets (as concept nodes represented by sets of synonyms are called in WordNet), which correspond each to a sense of the source word with a weighting based on its frequency of being used in that sense. Out of each SynSet, one has availability of any number of tokens synonymously used to convey the overall sense, and the frequency of the each token as a choice for expressing the sense. The approach we evaluate in this paper is to find the replacement token which maximizes the probability (and so the relative frequency) of the target word as a substitute (across all senses) for the source word. A competing alternative strategy to explore in this context is to adopt the replacement word which most closely approximates the relative frequency of the source word. The difference is that the maximized frequency word has a claim to greatest familiarity, and the balanced frequency substitution has the possibility of going unnoticed through its distributional equivalence to the source. Synonymy avoidance is a general cognitive constraint that people bring to lexical knowledge.[13] Lexical synonymy is extremely rare in natural language in part because people are cognitively driven to make semantic or pragmatic distinctions when candidate synonyms emerge. Thus, sharply contrasting frequency of a substitute for a source word may draw attention to a possible semantic distinction (such as typicality) that could go unnoticed if frequency is balanced. Nonetheless, here we evaluate the former approach, leaving the latter to future work.

Note that the WordNet transforms are not reliably reversible. For example, having replaced a word with a more general term in the hierarchy, for example, “car” with “vehicle”, it is not clear how to choose among specific kinds of vehicle (“bus”, “wagon”, etc.) to reverse the operation. The frequency effects on synonyms mentioned above also make those transforms generally non-reversible as well – for example the most likely synonymic replacement for “tariff” may be “duty”, but its most likely replacement is “responsibility”.

These three types of transforms have the advantage that they rely on part-of-speech tagging. Techniques that require deeper analysis, such as word sense disambiguation, syntactic parsing or the extraction of syntactic and semantic dependencies are more likely to fail because of lower accuracy of these technologies. Structural alternations are often not only structurally constrained, but involve complex semantic and discourse limitations,[14–16] although Murphy & Vogel[5] did identify a small number of predictable transforms.

Another advantage of the approach driven by lexical relations is linguistic portability. WordNet is available in many widely-spoken languages (including Arabic, Chinese, French, German, Hindi and Spanish) as are high quality POS taggers, though the scale and accuracy of these resources may vary. The methods are independent of particular fonts and scripts, in contrast to some extant approaches.[17, 18]

3. CORPUS BASED VALIDATION OF CANDIDATE MARKS

We validate the **structural fit** of a replacement against the original text. This may concern the grammatical fit between words (e.g. “this situation” → “this circumstances” instead of “these circumstances”; “long for” → “crave for” instead of “crave”), or conventional fit (e.g. “black and white” → “white and black”). This is achieved by testing the corpus frequency of the original phrase, versus the altered phrase, by looking at the number of matching documents returned for each by the Google search engine (a method that has been shown to correspond both to corpus frequencies and human plausibility judgements[19]). This constitutes a general corpus that is several orders of magnitude bigger than the 100 million word British National Corpus (BNC).[20] For example, in 2003 the Altavista search engine was estimated to contain 76 billion words of English.[21] The resulting relative frequency is a measure of the admissibility of the candidate change, relative to the original formulation.

In the case of adjective conjunction, we can look at **internal structural fit** by comparing the frequency of the original and altered ADJECTIVE CONJUNCTION ADJECTIVE phrases, (e.g. “lawful and unlawful” with “unlawful and lawful” in example (3)). We then look at the **external structural fit** by comparing document frequencies found for the left and right boundary word pairs (e.g. “between lawful” against “between unlawful”; and “unlawful acts” against “lawful acts”). There is no question of semantic fit/felicity, since no lexemes have been changed.

For closed-class category modifications like our relativiser switch, we only examine external structural fit, comparing e.g. the frequency of the phrases “women who are” with “women that are” (2). There is no internal structure to talk of, since it consists of only the relativiser. Since the word changed is a function word, it has no open semantics to interact surprisingly with surrounding words.

WordNet based substitutions do not meaningfully involve internal structure checks for the same reason as relativisers. External structural fit is checked by comparing the relative frequency of “picking hungrily at” and “picking ravenously at”

(8) (with independence assumptions, this is equivalent to checking the relative frequencies of left and right hand boundaries separately, and multiplying them). The **semantic fit** (or **coherence**) tests the relative frequency of *sets* of surrounding content words. This may concern stylistic and dialect fit (e.g. “lead violin in the Berlin Philharmonic” vs. “lead fiddle in the Berlin Philharmonic”), or conceptual fit (e.g. “she met her love on the banks of Lake Geneva” vs. “she met her love on the financial institutions of Lake Geneva”). For this validation we deem word inflections, function words, and ordering to be confounding factors, and so assess the number of documents returned by Google for the set of 5 nearest content word roots – for sentence (4) we compare {*quality, wine, superb, elegance, great*} against {*quality, wine, brilliant, elegance, great*}. This also allows us to widen the context window without running into the data sparseness problems associated with searching for exact phrases.

4. EVALUATION

While we apply an evaluation method to the techniques suggested in section 2, we stress that the method is intended as one that can be applied to test any candidate technique for performing undetectable meaning preserving text transforms.

4.1. Human and computer assessment

Topkara et al.[8] used evaluation metrics from Machine Translation to quantify the degree to which original and marked texts vary, as a measure of preservation of grammar, style and meaning. Since our method is based on lexical and ordering variations (while preserving meaning), such a metric is not appropriate here.

We think that human judgements of transform types are essential, since for all information applications, the aim is that perceived quality is not affected. However, human judgements are noisy, requiring large numbers of experimental items and subjects[22] to give reliable results. So, we attempt here to establish the relationship between *n*-gram type automatic measures of felicity[6, 12] and human judgements. The intent is that a baseline of automatically calculated measures related to human judgements can be reliably used to make predictions of items not directly tested by humans.

We applied our system (tokenisation, POS tagging, word stemming, candidate recognition, word infection, and replacement) to a random sample of 4000 sentences from the British National Corpus. 53 adjective swap and 160 relativiser switch candidates were found. For the WordNet based substitutions, applying a probability threshold of 0.8, and excluding candidates whose meaning was excessively general (e.g. “person”, “artifact”, “be”) by means of examining their depth and branching factors in the network, we found a total of 1036 candidate sentences, some with more than one possible transform.[‡] This resulted ca. 1250 candidate changes from which we chose 20 representative test items (5 adjective conjunction swap, 5 relativiser switch and 10 related word substitution), that were spread over the range of resulting felicity measures, and that included changes that were both successful and unsuccessful according to our own intuitions.

These items were presented to 25 subjects (colleagues and acquaintances of the authors) in arbitrary order, and preceded by two warm-up filler items. Subjects were first asked to judge how good they thought each altered sentence sounded on a seven-point scale, testing acceptability (plausibility, grammaticality and style). On a second page, the subjects were asked to compare the altered sentence with their originals, judging to what extent meaning was preserved, again on a seven-point scale. Other rating methods could have been employed and the nature of alternatives is the subject of a separate strand of our research, but this is an orthogonal issue to the general methodology we propose here. [§]

5. RESULTS

5.1. Analysis of individual items

It is interesting to consider what aspects of individual items elicited negative acceptability responses from humans, and also what factors interacted with their assessment that transforms did not preserve meaning (and where those judgements failed to correlate with the automated assessment). The set of sentences used in the judgement exercise are listed in the Appendix. For each test item measures of goodness are given, both human (acceptability and phrasal synonymy) and machine derived (internal/external structural fit, and semantic coherence) .

[‡] All sentences can be viewed at https://www.cs.tcd.ie/research_groups/clg/brian/markedBNCsample.html.

[§] The experiment was administered via the web. See https://www.cs.tcd.ie/research_groups/clg/brian/dummyIhSurvey1.html and .../dummyIhSurvey2.html for the precise format and instructions.

Table 1. Correlation of Human and Machine Measures

Human Judgement	Statistic	Internal Structure	External Structure	External Coherence
Acceptability	Pearson	.920	.468	.682
	2-tailed Sig.	**.027**	**.038**	**.030**
	N	5	20	10
Preservation of Meaning	Pearson	.914	.378	.311
	2-tailed Sig.	**.030**	.100	.381
	N	5	20	10

First of all, part of speech tagging errors resulted in unsuccessful transformation (for example (Appdx:5) and (Appdx:20) in the appendix where "very" and "final" were tagged as adjectives). Others are boundary errors such as "an welcome" in (Appdx:4)). Also automated productive morphology was too generative, see "fartherer" in (Appdx:19).

In several cases, people disagreed with the automated measures (all automated measures of relative frequency were log-normalized). For example, people were more tolerant of "that" as a relativiser with human referents ("lawyer" in (Appdx:9)), although this is actually very infrequent in corpora. People also tended to think the relative clause with the body "roamed around their new domain during the holidays" (Appdx:10) best analyzed as a non-restrictive relative due to plausibility with respect to modifying "his children", although because there was no signalling comma, the automated method concluded the alternation acceptable. The relative "that" is much less acceptable as a relativiser for nonrestrictive relative clauses than for restrictive relative clauses.[5, 15]

In two cases, the judgements were reasonable enough, but the automated techniques were biased by skews in the relative frequencies of the alternatives (for example phrases like "economic conditions" are much more frequent than "social conditions" (Appdx:2)). The outcome is that people thought the change meaning preserving, but the automated technique would have resisted approving it.

Finally, the conceptual measure based on the five-word window turned out to be too general to function as a good measure. See (Appdx:16) – the relevant window includes "save", "prosper", "vast", "unit" and "trust", yielding a low coherence value relative to the original window around "huge" by the automated method, while people (seemingly correctly) identified it as acceptable. The automated method fails because there really isn't a semantic connection among those words. Perhaps a larger window would improve performance for this technique.

5.2. Overall correlations

Results are summarized in table 1 (See the appendix for the average response by item). It can be seen that a significant correlation exists between grammatical acceptability as judged by humans and the automated predictors based on internal structure, external structure, and semantic coherence, calculated as in §3. Human judgements of meaning correlated most strongly with the internal structure. However, note that this measure only applies to five of the experimental items. Several variations on these machine measures were tried, including combining internal and external measures of structural fit, and normalising measures by relative frequencies of replacement words, but all of these combinations yielded weaker associations with human judgements.

The results summarized express the effects of the factors in isolation. A regression model for the data can capture interactions. Note that certain interactions are nearly analytic: unacceptable sentences were uniformly assessed as not having preserved meaning under transformation.

The regression model is given as follows, where "A" is human judgement of acceptability; "M" meaning-preservation; "IS" is internal structural rating; "ES" measures external structure; "C" represents external coherence.

(9)
$$A = 1.685 + 0.585 * IS + 0.065 * ES + 0.122 * C$$

(10)
$$M = 1.466 + 0.184 * IS + 0.094 * ES + 0.067 * C$$

The regression model for acceptability (9) is highly correlated (Pearson's coefficient of 0.795) and significant ($p = 0.001$). The regression model for phrasal synonymy, or meaning preservation, (10) is less strongly correlated (Pearson's coefficient of 0.422), but this result is not significant ($p = 0.356$). In both cases the measure of internal structure (*IS*) provides the bulk of the predictive power, while for acceptability semantic coherence is surprisingly the second strongest term.

The outcome of the regression model is that we are able to predict for a novel test item, using the suggested automatic assessment methods, what a human judgement of acceptability is likely to be. But the methods evidently do not allow us to predict human assessment of synonymy between the source and transformed items with a high degree of confidence, thus, compromising the meaning preservation properties of the suggested transforms. Nonetheless, we examine what those predictions would be.

Recalling that for acceptability we offered a scale ranging from -3 to +3 with -3 as completely unacceptable, and the same range for phrasal synonymy with +3 indicating absolute meaning identity, we identified threshold values appropriate to each dimension. Based on our independent assessment of the lowest average rating that we agreed with and with reference to similar levels in a previous similar study,[5] a rating of 1 was our cut-off point for acceptability and 0.88 for phrasal synonymy, exemplified by (Appdx:21) and (Appdx:18), respectively.

Applying the models to the entire random sample of 4000 sentences, we find that 24% are predicted to exceed the acceptability threshold and 24% are expected exceed the phrasal synonymy threshold, and 19% of the sentences are anticipated to pass both. Given an average sentence length of 20 words, that means that we estimate one transform to be safely applicable for every 100 words.

6. CONCLUSIONS

In light of the poor performance with respect to meaning preservation, our next step in this research is to explore the alternative computation suggested in §2 of making substitutions that preserve rather than maximize relative frequency. Broadening the window of words used for the coherence measure might also have a beneficial effect, as long as a back-off remained available in case of data sparseness. In the small sample of sentences used for human evaluation, hyponym candidates (11, 12, 13) from WordNet performed uniformly badly, so it need to be investigated if this is a general problem with the approach. More generally, if larger numbers of web queries were available (the Google API limits a user to 1000 queries per day), the thresholds from the WordNet module could be lowered to allow more candidates through for web validation. This would slow the system down, but things could be speeded up somewhat by precompiling WordNet replacement candidate lists for all 150,000 entries it contains.

Other sorts of shallow transform types also need to be explored, for example the equivalence of double negations and antonyms (11).

(11) The main points of contention, which are ~~not insurmountable~~ surmountable, seem to be the role of monarchy and disarmament.[¶]

Feedback from experiment subjects indicated that for some sentences that they found unacceptable, the reason was not related to the automatic transformation made. More sophisticated paradigms, such as eye-tracking experiments, might allow us to localise the source of low grammaticality judgements, and so remove one confounding factor.

While the result is partly positive and partly negative, it is sufficiently encouraging to pursue further, given that we have obtained some insights into the effects of frequency in the data on human assessments of acceptability and meaning preservation.

Acknowledgements

We gratefully acknowledge the support of Science Foundation Ireland funding in the Research Frontiers Programme project 05/RF/CMS002.

[¶] http://archive.gulfnews.com/articles/06/08/09/10058587.html, last verified August 15, 2006.

REFERENCES

1. K. Winstein, “Lexical steganography through adaptive modulation of the word choice hash.” http://www.imsa.edu/ keithw/tlex/, 1998.
2. M. Atallah, V. Raskin, M. Crogan, C. Hempelmann, and K. Kerschbaum, “Natural language watermarking: Design and analysis and a proof-of-concept implementation,” in *Proceedings of Forth International Workshop on Information Hiding*, 2001.
3. B. Murphy, “Syntactic information hiding in plain text,” Master’s thesis, Trinity College Dublin, https://www.cs.tcd.ie/research_groups/clg/brian/murphy01hidingMasters.pdf, September 2001.
4. M. Atallah, V. Raskin, C. Hempelmann, M. Karahan, R. Sion, and K. Triezenberg, “Natural language watermarking and tamperproofing,” *Lecture Notes in Computer Science and Proceedings of the 5th International Information Hiding Workshop* , 2002.
5. B. Murphy and C. Vogel, “The syntax of concealment: Reliable methods for plain text information hiding,” in *Proceedings of Security and Steganography and and Watermarking of Multimedia Contents IX*, (San José), January 2007.
6. C. M. Taskiran, U. Topkara, M. Topkarab, and E. J. Delp, “Attacks on lexical natural language steganography systems,” in *Proceedings of the SPIE International Conference on Security and Steganography, and Watermarking of Multimedia Contents*, (San Jose), January 2006.
7. Y. Miyao and J. Tsujii, “Probabilistic disambiguation models for wide-coverage HPSG parsing,” in *Proceedings of of 43rd ACL*, pp. 83–90, 2005.
8. M. Topkara, G. Riccardi, D. Hakkani-Tür, and M. J. Atallah, “Natural language watermarking: Challenges in building a practical system,” in *Proceedings of the SPIE International Conference on Security and Steganography, and Watermarking of Multimedia Contents*, (San Jose), January 2006.
9. E. Brill, “A simple rule-based part-of-speech tagger,” in *Proceedings of ANLP-92 and 3rd Conference on Applied Natural Language Processing*, pp. 152–155, (Trento), 1992.
10. E. Brill, “Some advances in transformation-based part of speech tagging,” in *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, (Seattle), 1994.
11. G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, , and K. Miller, “Introduction to WordNet: an on-line lexical database,” *International Journal of Lexicography* **3**(4), pp. 235–244, 1990.
12. I. A. Bolshakov, “A method of linguistic steganography based on collocationally-verified synonymy,” in *Information Hiding and 6th International Workshop*, J. J. Fridrich, ed., *Lecture Notes in Computer Science* **3200**, pp. 180–191, Springer, Berlin, 2004.
13. A. Carstairs-McCarthy, “Synonymy avoidance and phonology and the origin of syntax,” in *Approaches to the Evolution of Language: Social and Cognitive Bases*, J. R. Hurford, M. Studdert-Kennedy, and C. Knight, eds., Cambridge University Press, Cambridge, 1998.
14. S. R. Anderson, “On the role of deep structure in semantic interpretation,” *Foundations of Language* (7), pp. 387–396, 1971.
15. D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, *Longman grammar of spoken and written English*, Longman, Harlow, 1999.
16. B. Murphy and C. Vogel, “Cross-linguistic empirical analysis of constraints on passive,” in *Presentation to the Symposium on Interdisciplinary Themes in Cognitive Language Research*, (Helsinki), 2005.
17. J. Brassil, S. H. Low, N. F. Maxemchuk, and L. O’Gorman, “Electronic marking and identification techniques to discourage document copying,” *IEEE Journal on Selected Areas in Communications* **13**(8), pp. 1495–1504, 1995.
18. Sun Xingming, Luo Gang, and Huang Huajun, “Component-based digital watermarking of chinese texts,” in *Proceedings of the 3rd International Conference on Information Security*, (Shanghai), November 2004.
19. F. Keller and M. Lapata, “Using the web to obtain frequencies for unseen bigrams,” *Computational Linguistics* **29**(3), pp. 459–484, 2003.
20. G. Burnage and D. Dunlop, “Encoding the british national corpus,” in *English Language Corpora: Design and Analysis and Exploitation: Papers from the 13th international conference on English Language research on computerized corpora*, J. Aarts, P. de Haan, and N. Oostdijk, eds., Rodopi Press, (Amsterdam), 1992.
21. A. Kilgarriff and G. Grefenstette, “Introduction to the special issue on the web as corpus,” *Computational Linguistics* , 2003.
22. C. T. Schütze, *The Empirical Basis of Linguistics*, University of Chicago Press, Chicago, 1996.

Appendix

Experiment Test Items and Average Responses

1. Efforts to enhance labour mobility and to remove **fiscal and legal obstacles** are necessary conditions for the success of this scenario.

   Acceptability: 2; Phrasal Synonymy: 2.92; Internal Structure: -0.83; External Structure: 1.48; Coherence: n.a.

2. That the appropriate award for non-pecuniary loss may vary in differing **economic and social conditions** in different parts of the world has been stated by the Privy Council in Jag Singh v Toong Fong Omnibus Co[1964] 1 WLR 1382, per Lord Morris:

   Accept: 1.88; PhrSyn: 2.75; Int. Struct: 0.36; Ext. Struct: -0.56; Coher: n.a.

3. I am deeply honoured to be invited to this **lavish and momentous occasion** by my esteemed friends, Martha and George.

   Accept: 2.41; PhrSyn: 2.5; Int. Struct: 0; Ext. Struct: 5.53; Coher: n.a.

4. An additional plus for the Anatom is its lining, which has in-built Antibac, an antibacterial system which slows down the growth of bacteria on the skin, an **welcome but unusual concept** in shoe technology.

   Accept: 0.81; PhrSyn: 1.17; Int. Struct: -3.16; Ext. Struct: -2.84; Coher: n.a.

5. He was very **very and courteous kind**.

   Accept: -1.96; PhrSyn: -0.54; Int. Struct: -4.67; Ext. Struct: -0.2; Coher: n.a.

6. There is one final item **that (which)** appears valuable but innocuous enough, if one isn’t a daemonologist.

   Accept: 1.48; PhrSyn: 1.75; Int. Struct: n.a.; Ext. Struct: 1.47; Coher: n.a.

7. We even had a walnut tree **that (which)** is still going, but it does not bear fruit any more.

   Accept: 1.85; PhrSyn: 1.71; Int. Struct: n.a.; Ext. Struct: 0.25; Coher: n.a.

8. Physics is interested only in those abstracted features of the world **that (which)** its theories specify: one way of describing what physics does is “go beneath” how the world appears to us to uncover the “real” physical principles and processes **that (which)** produce the ordered universe.

   Accept: 2.04; PhrSyn: 1.63; Int. Struct: n.a.; Ext. Struct: 5.25; Coher: n.a.

9. Both were boys of 8 years old and had their imaginations fired by Perry Mason, a fictional TV lawyer **that (who)** helped the poor.

   Accept: 1.3; PhrSyn: 1.54; Int. Struct: n.a.; Ext. Struct: -5.91; Coher: n.a.

10. This did not prevent him paying for the installation of a swimming pool to amuse his children **that (who)** roamed around their new domain during the holidays.

    Accept: 1.15; PhrSyn: 0.83; Int. Struct: n.a.; Ext. Struct: -1.66; Coher: n.a.

11. Do you want Dovelands put back ? ... cos Dovelands didn’t come on this year either and this is a **double-bind (dilemma)** that will face us all.

    Accept: 0.15; PhrSyn: 0.17; Int. Struct: n.a.; Ext. Struct: -6.43; Coher: -1.41

12. The **massagers (physiotherapists)** ... will come round and they come round regularly and exercise your muscles

    Accept: 0.69; PhrSyn: -1.08; Int. Struct: n.a.; Ext. Struct: -4.47; Coher: 2.02

13. A prospective **freeloader (borrower)** would approach a number of investment banks, one of which would be invited to act as the lead manager of an issue.

    Accept: 2.22; PhrSyn: -1.33; Int. Struct: n.a.; Ext. Struct: -4.29; Coher: -4.53

14. With new firms in the North of England under cutting their city **competitors (rivals)**, is the writing on the wall for one of London’s most traditional professions.

    Accept: 2.46; PhrSyn: 2.25; Int. Struct: n.a.; Ext. Struct: -1.73; Coher: 2.26

15. The hubbub outside comes not from the picturesque **dealers (traders)** of the bazaar but from some 500 fans pleading for a **glance (glimpse)** of their idol.

    Accept: 2.35; PhrSyn: 1.96; Int. Struct: n.a.; Ext. Struct: 0; Coher: -0.04

16. The China Dragon Fund is managed by Save & Prosper, the **vast (huge)** unit trust company, so it has a wealth of experience behind it.

    Accept: 2.33; PhrSyn: 1.79; Int. Struct: n.a.; Ext. Struct: -1.21; Coher: -0.43

17. If there is one worry about it, it is that the national **program (curriculum)** and the attainment targets will only partly tell you about a child.

    Accept: 1.15; PhrSyn: 1.08; Int. Struct: n.a.; Ext. Struct: -1.78; Coher: 0.89

18. I think we have gone in at a very **reasonable (sensible)** price, and I am sure the other contractors did.

    Accept: 1.19; PhrSyn: 0.88; Int. Struct: n.a.; Ext. Struct: 8.29; Coher: 1.43

19. Those staying for a **fartherer (further)** week will be transferred directly from Kufstein to their resort.

    Accept: -1.44; PhrSyn: -0.5; Int. Struct: n.a.; Ext. Struct: -11.82; Coher: -15.01

20. He received a walkover into the **last (final)** when Richard Krajicek, who may well be the next known to win Wimbledon, withdrew with tendinitis in his right shoulder.

    Accept: 1.77; PhrSyn: -0.88; Int. Struct: n.a.; Ext. Struct: 0.97; Coher: 0.38

21. FILLER: You can easily opt-out at any time, to stop any further **contact (communication)** from us.

    Accept: 1.19; PhrSyn: 2

22. FILLER: Throw **the idol me (me the idol)**

    Accept: -1.3; PhrSyn: -0.39