Modelling semantic transparency in English compound nouns

Melanie J. Bell
Anglia Ruskin University
Cambridge
U.K.
melanie.bell@anglia.ac.uk

Martin Schäfer
Friedrich Schiller University
Jena
Germany
post@martinschaefer.info

1 Introduction

Semantic transparency is known to play an important role in the storage and processing of complex words (e.g. Marslen-Wilson et al. 1994), and human raters of transparency achieve high levels of agreement (e.g. Frisson et al. 2008, Munro et al. 2010). In the case of noun-noun compounds, overall transparency is largely determined by the transparency of the individual constituents. For example, Reddy et al. (2011) showed that the perceived transparency of a compound is highly correlated with both the sum and the product of the perceived transparencies of its constituents. Furthermore, many psycholinguistic studies find significant effects for semantic transparency using a four-way distinction based on perceived constituent transparency: transparent-transparent (e.g. *carwash*), transparent-opaque (e.g. *jailbird*), opaque-transparent (e.g. *strawberry*) and opaque-opaque (e.g. *hogwash*) (Libben et al. 2003). Bell and Schäfer (2013) modelled the transparency of individual compound constituents and showed that shifted word senses reduce perceived transparency, while certain semantic relations between constituents increase it. However, this finding is problematic in at least two ways. Firstly, it is not clear whether there is a solid basis for establishing whether a specific word sense is shifted or not. For example, *card* in *credit card* is clearly shifted if viewed etymologically, but may not synchronically be perceived as shifted due to its frequent use. Secondly, work on conceptual combination by Gagné and collaborators has shown that relational information in compounds is accessed via the concepts associated with individual modifiers and heads, rather than independently of them (e.g. Spalding et al. 2010 for an overview). This leads to the hypothesis that it is not whether a specific word sense is etymologically shifted, nor whether a specific semantic relation is used *per se*, that makes a compound constituent more or less transparent; rather, it is the degree of expectedness of a particular word sense and a particular relation for a given constituent. In this paper, we provide evidence in support of this hypothesis: the more expected the word sense and relation for a constituent, the more transparent it is perceived to be.

2 Method

We used the publicly available dataset described in Reddy et al. (2011), which gives human transparency ratings for a set of 90 compound types and their constituents (N1 and N2), and comprises a total of 7717 ratings. To model the expectedness of word senses and semantic relations for a given compound constituent, we used the constituent families of the compounds, which we extracted in a two step process. We took all strings of exactly two nouns that follow an article in the British National Corpus and which also occur four times or more in the USENET corpus (Shaoul and Westbury 2010). From this set, we extracted the positional constituent families for all constituent nouns in the Reddy et al. dataset, giving a total of 4553 compounds for the N1 families and 9226 for the N2 families. Each of these compound types was coded for the semantic relation between the constituents (after Levi 1978), and for the WordNet sense of the constituent under consideration (Princeton 2010). We then calculated the proportion of compound types in each constituent family with each semantic relation (relation proportion), and each WordNet sense of the constituent in question (synset proportion). We take these two measures to reflect the expectedness of the respective relations and WordNet senses of the constituents: if a relation or sense occurs in a high proportion of the constituent family, it is more expected. These variables were used, along with other quantitative measures, as predictors in ordinary least squares regression models of perceived constituent transparency. The final model for the transparency of N1 is given in Table 1:

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

	Coef	S.E.	t	Pr(>|t|)
Intercept	-4.6413	0.6593	-7.04	<0.0001
relation proportion in N1family	-0.2187	0.6013	-0.36	0.7161
log family size of N1	-0.0189	0.0931	-0.20	0.8395
synset proportion in N1family	-0.2426	0.6152	-0.39	0.6934
log synset count of N1	-0.7939	0.2469	-3.22	0.0013
compound proportion in N1 family (token-based)	3.0130	0.6788	4.44	<0.0001
log frequency of N1	0.8728	0.0569	15.34	<0.0001
relation proportion * log family size	0.3311	0.1305	2.54	0.0113
synset proportion * log synset count	0.6855	0.3161	2.17	0.0303
compound proportion * log frequency N1	-0.2804	0.0816	-3.44	0.0006

Table 1: Final model for the transparency of N1, R^2 adjusted = 0.334

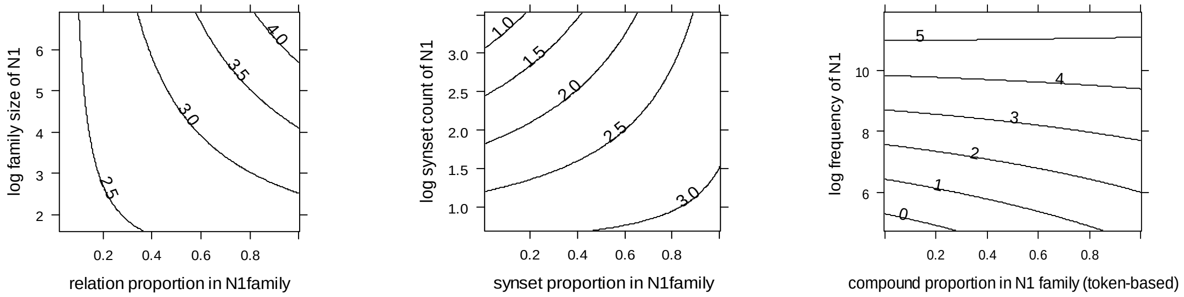

Figure 1. Interaction plots for N1 transparency

3 Results

All predictors in our model enter into significant interactions, and these are shown graphically in Figure 1, where the contour lines on the plots represent perceived transparency of the first constituent (N1). The first plot shows an interaction between relation proportion and overall (log) family size: for small families, relation proportion plays little role, whereas for larger families, in accordance with our hypothesis, the transparency of N1 increases with the proportion of the corresponding relation in the family. The second plot shows the interaction between the synset proportion and the total number of a constituent's senses (as listed in WordNet): only if there is a sufficient number of different senses in the family is their proportion a reliable predictor of semantic transparency. There is also a small but significant interaction between the log frequency of a constituent and the proportion of the constituent family (in terms of tokens) represented by the compound in question: this shows that transparency increases with frequency, but only in the lower frequently ranges does the proportion in the family play a role.

4 Conclusion

Overall, the model provides clear evidence for our hypothesis. N1 is rated as most transparent when it is a frequent word, with a large family, occurring with its preferred semantic relation and most frequent sense, and with few other senses to compete. We interpret the results as indicating that compound constituents are perceived as more transparent when they are more expected (both generally and with a specific sense) and when they occur in their most expected semantic environments. In information theory, the less expected an event, the greater its information content: in so far as perceived transparency is a reflection of expectedness, it can therefore also be seen as the inverse of informativity.

Acknowledgements

This work was made possible by three short visit grants from the European Science Foundation through NETWORDS - The European Network on Word Structure (grants 4677, 6520 and 7027), for which the authors are extremely grateful.

References

Bell, Melanie J. and Martin Schäfer. 2013. Semantic transparency: challenges for distributional semantics. In Aurelie Herbelot, Roberto Zamparelli and Gemma Boleda eds., *Proceedings of the IWCS 2013 workshop: Towards a formal distributional semantics*, 1–10. Potsdam: Association for Computational Linguistics.

Frisson, Steven, Elizabeth Niswander-Klement and Alexander Pollatsek. 2008. The role of semantic transparency in the processing of English compound words. *British Journal of Psychology* 991, 87–107.

Levi, Judith N. 1978. *The syntax and semantics of complex nominals*. New York: Academic Press.

Marslen-Wilson, William, Lorraine K. Tyler, Rachelle Waksler and Lianne Older. 1994. Morphology and meaning in the English mental lexicon. *Psychological Review* 101, 1: 3-33.

Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai , Robin Melnick, Christopher Potts, Tyler Schnoebelen and Harry Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 122-130. Association for Computational Linguistics.

Princeton University. 2010. *WordNet*.
<http://wordnet.princeton.edu>

Reddy, Siva, Diana McCarthy and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of The 5th International Joint Conference on Natural Language Processing 2011 IJCNLP 2011*, Chiang Mai, Thailand

Shaoul, Cyrus and Chris Westbury. 2010. An anonymized multi-billion word USENET corpus 2005-2010
http://www.psych.ualberta.ca/˜westburylab/downloads/usenet.download.html

Spalding, Thomas L., Christina L. Gagné, Allison C. Mullaly and Hongbo Ji. 2010. Relation-based interpretation of noun-noun phrases: A new theoretical approach. *Linguistische Berichte Sonderheft* 17, 283-315

Wurm, Lee H. 1997. Auditory processing of prefixed English words is both continuous and decompositional. *Journal of Memory and Language*, 37, 438–461.