Research Networking Programmes

Short Visit Grant ☒ or Exchange Visit Grant ☐

Scientific Report

***Proposal Title**:* Semantic detail in the computation of compound meanings and its influence on semantic transparency

Application Reference N°: 6520

1) Purpose of the visit

Bell & Schäfer (2013) modelled the transparency both of compound words and 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 led us 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. The aims of this short visit between Martin Schäfer, the visitor, and Melanie Bell, the host, were to test this hypothesis and to write up the findings for journal submission.

2) Description of the work carried out during the visit

We used the publicly available dataset described in Reddy et al (2011), which gives human transparency ratings for a set of compounds and their constituents (N1 and N2). 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 & Westbury 2010). From this set, we extracted the positional constituent families for all noun constituents in the Reddy et al dataset. Every compound type represented in the family of any left-hand constituent (N1) was then coded for the semantic relation between the constituents (after Levi 1978), and for the WordNet sense

of that constituent (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 constituent transparency.

3) Description of the main results obtained

The final model for the transparency of N1 is given in Table 1:

	Coef	S.E.	t	Pr(>|t|)
Intercept	-4.6413	0.6593	-7.04	<0.0001
relPropInN1Fam	-0.2187	0.6013	-0.36	0.7161
logFamSizeN1	-0.0189	0.0931	-0.20	0.8395
synsetPropInN1Fam	-0.2426	0.6152	-0.39	0.6934
logSynsetCountN1	-0.7939	0.2469	-3.22	0.0013
compoundTokenPropInN1Fam	3.0130	0.6788	4.44	<0.0001
logFreqN1	0.8728	0.0569	15.34	<0.0001
relPropInN1Fam * logFamSizeN1	0.3311	0.1305	2.54	0.0113
synsetPropInN1Fam * logSynsetCountN1	0.6855	0.3161	2.17	0.0303
compoundTokenPropInN1Fam * logFreqN1	-0.2804	0.0816	-3.44	0.0006

Table 1: Final model for the transparency of N1, R2 adjusted = 0.334

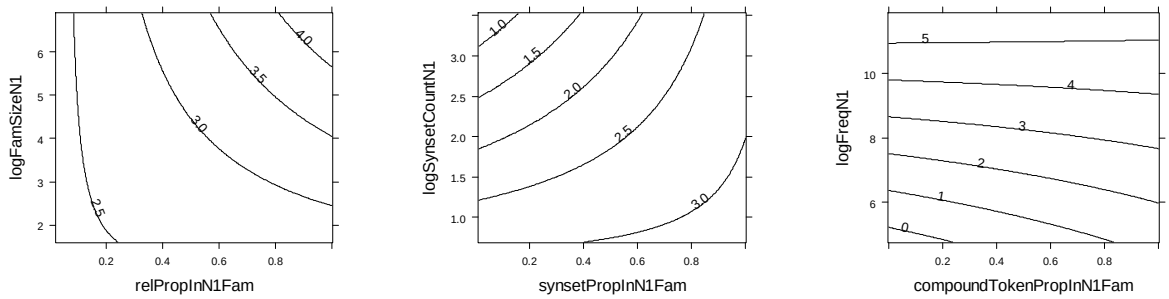

Figure 1: Interaction plots for N1 transparency

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 left-hand constituent (N1). The first plot shows an interaction between relation proportion and overall (log) family size. In general, the transparency of N1 increases with the proportion of the corresponding relation in the family, but this effect is greatest for constituents with large positional families. Perceived transparency is greatest when N1 has a large family, and the compound has the dominant relation for that family. The second plot shows the interaction between the synset proportion and the total number of a constituent's senses (as listed in WordNet): overall, perceived transparency increases with synset proportion, and this effect is greatest for constituents with a large number of different senses in their family. Perceived transparency is lowest when N1 has a large number of senses and the compound involves a sense that occurs rarely. 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.

Overall, the model provides clear evidence for our hypothesis. N1 is most transparent when it is most expected: 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. 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.

4) Future collaboration with host institution (if applicable)

We have submitted an abstract for a journal article, reporting our findings, to a planned special issue of *Morphology*. By the end of September 2014, we will have finished coding the N2 constituent families, so that we will be able to include this data in our final model, as well as the N1 data reported here. We will also present the results of our new work in joint talks at research institutes and conferences, e.g. at this year's Linguistics Association of Great Britain meeting at the University of Oxford (1-5.09.2014).

In the longer term, we plan to extend the work to investigate contextual effects in the interpretation of compounds.

5) Projected publications / articles resulting or to result from the grant *(ESF must be acknowledged in publications resulting from the grantee's work in relation with the grant)*

Schäfer, Martin & Melanie J. Bell. Modelling semantic transparency. *Morphology*.

6) Other comments (if any)

We are extremely grateful for the funding through Networds – The European Network on Word Structure. This is the second time that a short visit grant has allowed Martin Schäfer to spend two weeks in Cambridge, and this has really benefitted us. Both the fruitful face-to-face discussions, and also the blocking of time for one dedicated purpose, have led to important advances in our work and our thinking. We would therefore welcome any opportunity to contribute and give back to this network in the future.

References

Bell, Melanie J. & Martin Schäfer. 2013. Semantic transparency: challenges for distributional semantics. In Aurelie Herbelot, Roberto Zamparelli & Gemma Boleda (eds.), Proceedings of the IWCS 2013 workshop: Towards a formal distributional semantics, 1–10. Potsdam: Association for Computational Linguistics.
URL: http://www.aclweb.org/anthology/W13-0601

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

Princeton University. 2010. About WordNet. http://wordnet.princeton.edu

Reddy, Siva, Diana McCarthy, Suresh Manandhar. 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

All data for the paper is available from the following site: http://sivareddy.in/papers/files/ijcnlp_compositionality_data.tgz.

Shaoul, Cyrus & 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. In Olsen, Susan, ed., New Impulses in Word-Formation, *Linguistische Berichte* Sonderheft 17, 283-315.