Scientific Report
Short visit grant, ESF-activity *European Network on Word Structure: Cross-disciplinary Approaches to Understanding Word Structure in the Languages of Europe*
Reference Number 4677

Applicant: Martin Schäfer

May 7, 2012

1 Purpose of the visit

The purpose of the collaboration between Martin Schäfer, the applicant, and Melanie Bell, the host, is, firstly, to work towards a clarification of the notion of semantic transparency and its status in the analysis of complex words, and, secondly, to operationalise this notion by identifying its empirical correlates.

The immediate aims of the visit as stated in the application were a) to formulate hypotheses and methodology for our empirical work and b) to write a full-scale funding proposal. However, in our consultations before the visit, it became apparent that the second goal could not in fact be achieved (see below), and we decided instead to dedicate our time to the first goal.

The reason for not writing a full-scale project proposal within the two weeks of the stay was that it took some time to find a grant-giving scheme that might fulfil our requirements. Before the visit, we spoke to the funding experts at our two institutions (that is, the Servicezentrum Forschung und Transfer of the Friedrich-Schiller-Universität Jena and the Research Development and Commercial Service of Anglia Ruskin University), and formed the impression that there was no grant-scheme that would immediately suit our needs (i.e., that would involve two partners from Germany and the UK and would allow us to carry out experiments and other empirical research and finance mutual visits). Instead, the available schemes were either only for exchange visits without funding for empirical work or for larger scale projects, e.g. EU-programmes involving at least a third partner from a third country. However, we continued our enquiries during the two weeks and, after some delay in being able to contact the person responsible, eventually identified the 'key issues in the humanities' program of the Volkswagen Stiftung as a possible source of funding. We are currently working on a proposal for this program (the deadline is the first of June, 2012).

2 Description of the work carried out during the visit

As mentioned in the previous section, the aim we set ourselves for the visit was a clarification of the notion of semantic transparency, as well as the development of hypotheses for further empirical work on the issue.

Since the clarification involved a thorough reading of the previous literature, we decided that it would be best to channel that into the writing of a comprehensive review article on the issue. The main work carried out during the visit was thus geared towards writing this review article. This quickly lead to very fruitful exchanges with regard to a large number of conceptual issues involved in the notion of semantic transparency.

Coming from two different backgrounds ourselves, empirical language research on the one hand and formal semantics on the other hand, we first had to reach a mutual understanding about our own terminological conventions. In addition, the usage of the term *semantic transparency* in the literature is by no means uniform, ranging from a very basic definition used in psycholinguistic literature to a variety of usages of the term in more general morphological literature. Furthermore, different research traditions, sometimes tied to specific languages, relate the term to a variety of other concepts, including institutionalization, lexicalization, semantic compositionality and analyzability.

What we ended up doing was to develop a simple model that allowed us to pinpoint specific factors that are responsible for the semantic transparency or opacity of a construction, and to develop first hypotheses from there. This will be related in more detail in the next section.

3 Main results obtained

This section gives a brief overview of our main results, starting with an outline of the phenomenon itself and ending with a first few hypotheses to investigate in future work.

3.1 Outline of the phenomenon of semantic transparency

As far as the domain of our investigation is concerned, we restricted ourselves to noun noun and adjective noun combinations that are usually considered to be compounds, as well as any noun noun or adjective noun combination that can occur in the same position, i.e., that fits the slot after the determiner in a noun phrase ([det __]$_{\text{NP}}$). As far as languages are concerned, we restricted ourselves to English, Dutch, and German. In the following two sections, we first show the range of combinatorial possibilities that has to be considered, and secondly discuss a few terms that frequently play a role in the discussion of semantic transparency. For ease of exposition, we furthermore restrict ourselves to binary pairings, using A for the first part and B for the second part of the pairing.

3.1.1 The possibility space

In the psycholinguistics literature, e.g. Libben, Gibson, Yoon & Sandra (2003), one typically finds classifications of A B combinations into four groups, opaque-opaque, opaque-transparent, transparent-opaque, and transparent-transparent, corresponding to

English examples like *blackguard, strawberry, jailbird* and *schoolteacher* respectively. Two diagnostics that are regularly used in establishing this kind of classification are a) the extent to which the meaning of AB is predictable from the meanings of A and B individually, and b) the extent to which A and B retain the meanings they have in isolation when they become part of the complex construction. Thus, the reason for considering a word like *jailbird* as a transparent-opaque combination is that the meaning of *jail* is somehow related to the meaning of the whole (a jailbird is somebody who is frequently incarcerated in jail), and the meaning of *jail* here corresponds to its usage on its own. In contrast, a jailbird is not a bird, and the meaning of *bird* employed here does not correspond to its usage as a stand-alone item.

However, closer examination reveals that the situation is more complex, and that the fourfold distinction does not adequately capture this complexity. *Jailbird* is a good example to show this. While it is true that a jailbird is not a bird in the biological sense, it is also true that a word like *bird* can be used metaphorically to refer to persons, as in *He's a strange bird.* Once we have this metaphorical reading, the whole construction is very regular.

The scheme in figure 1 is meant to capture these complex possibilities. First of all, we assume that an underspecified relation R links the A and the B members of the construction. Secondly, we assume that the A as well as the B part can be shifted from their literal meaning to a metaphorical or metonymical reading. However, even after a shift, they are still linked to the other part of the construction via the R relation. Thirdly, we assume that context and world knowledge play a key role in establishing the meaning of a specific A B combination: they specify the relation R, and initiate the shifts of A and/or B. Note that the context for any given element in this scheme includes all other factors in the diagram, e.g., the context that initiates a shift for A includes the current setting for the relation R as well as B or its shifted variant, B'.

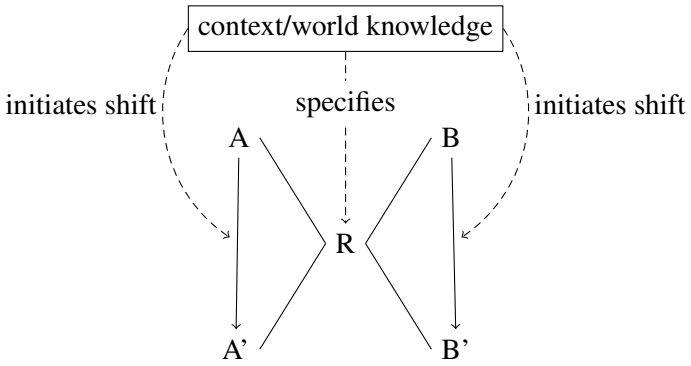

Figure 1: Scheme for A B combinatorics

Some examples that illustrate the general idea are discussed in the following. Thus, for a copulative compound like *singer-songwriter*, the property of being a singer and the property of being a songwriter are assigned to the very same individual; the relation R is set to identity. In this case, neither A nor B are shifted. If we keep A and B unshifted, we can vary the relation R. It can, for example, express a locative relation: a *country town* is a town that is located in the country. Here, R links the entity x that is a town to the entity y that is the country by providing the localization. While this case

illustrates a rather basic type of localization, R can be specified in much more detail. Thus, a *schoolteacher* is not a teacher who is located at a school, but a teacher who is trained to work for a school and/or someone who works as a teacher in a school. The corresponding R relation is thus *x who is trained to work for and/or who works at y*. Turning to shifted A and Bs, *jailbird* already illustrated the possibility of a shifted B in combination with an unshifted A, that is, an AB' combination. In this specific case, the relation R is specified to *x is regularly incarcerated in y*. An example for an A'B combination, that is, a shifted A combined with an unshifted B, is a combination like *stone t-shirt*: while *t-shirt* retains its literal meaning, *stone* is shifted to *resembling the colour of a stone*. The relation R is specified to *x has the colour y*. Finally, *buttercup* is an example of a compound where both A and B are shifted, that is, an A'B' pairing: *cup* is shifted to *flower with parts having a cup-shape*, and *butter* is shifted to *colour that resembles the colour of butter*. As this last example shows, the shifts involved may massively differ in quality, cf. especially the relatively easy shift in *bird* for person vs. the shift involved in *cup* for flower with parts in cup-shape.

Note that although all of the examples so far have been drawn from the domain of compounds, the processes involved are not specific to compounds. Thus, German adjectives for materials can easily be combined with common nouns like *Löwe* 'lion', e.g. *bronzener Löwe* 'bronze lion'. In this phrasal adjective-noun combination B is shifted to B': something resembling a lion in shape. The same adjective in combination with a noun like *Haut* 'skin', however, is shifted from material to colour of the material, yielding an A'B combination: *bronzene Haut* 'bronze skin'. And finally, the many lexicalized adjective-noun phrases in Dutch and German contain ample illustrations of A'B' combination, e.g. *kalter Kaffee* 'cold coffee', i.e., old news.

3.1.2 Related terms

Some of the terms and concepts that play a role in the discussion of semantic transparency have already been mentioned. For example, lexicalization plays a role, and the example of *buttercup* discussed in the previous section illustrates this quite clearly: a lexeme of a language may develop so idiosyncratically that the meaning shifts involved can only be understood with hindsight. Institutionalization is another concept that plays a role. Usually, it is applied to constructions where the meaning has been narrowed down to a more or less arbitrary definition, e.g., the German compound *Großstadt* 'big town' is, in the official terminology used in legislation and in administrative contexts, used only for towns with at least 100,000 inhabitants. Another relevant notion, analyzability, is sometimes used to mean that the parts make a discernible semantic contribution to the meaning of the whole, and it can thus be seen as the lower end of semantic transparency for complex constructions. Thus, *blackguard* in its oral form (/ˈblagɑːd/ or /ˈblægərd/ in the UK and the US, respectively) is not analyzable, because the fact that it was originally a compound is obscured in the synchronic phonology. The final two terms that need to be mentioned here are on the one hand *semantic compositionality*, and on the other hand *phrase-like semantics*. As far as we can see, *semantic compositionality* occurs with a very wide range of usages. It is sometimes used for any case where we can build up a complex semantic structure according to clear rules, even if these rules contain underspecifications or anchor points for pragmatics. In other cases, semantic compositionality is equated

with semantic transparency. The final term, *phrase-like semantics*, is often employed in the discussion of compounds and is usually meant to indicate that they are semantically very transparent. However, as our examples at the end of the previous section have shown, AB phrases can also be semantically very intransparent. Without further specification, this term is thus only of limited use.

3.2 First hypotheses

We hypothesise that semantic transparency can best be understood as a graded phenomenon, and that it will be possible to model this phenomenon in terms of distributional criteria. Furthermore, we hypothesise that the degree of semantic transparency, and therefore its distributional correlates, will be predictive for phenomena sometimes taken to mark a distinction between morphological and syntactic objects, including stress in English NN combinations, and the availability of AB constituents for anaphoric reference, coordination and modification.

With regard to semantic transparency, we hypothesise that the key to understanding this notion lies in a rigorous disentanglement of the different factors that we identified in section 3.1.1. Thus, it is not enough to consider wholesale notions like the four pairings opaque-opaque, opaque-transparent, transparent-opaque, and transparent-transparent. Instead, we must identify the exact relation R that holds between the two elements, and the extent to which the elements themselves have been shifted from their literal meaning.

However, not only do we assume that we need at least to distinguish these different factors, we also believe that all of these factors come with scaled costs. That is, it makes a difference with regard to semantic transparency whether the relation R is resolved to a very universal and simple notion (e.g. identity or localization) or to something more specific (e.g. as in *oil sand* 'intermixed with quantities of'). Likewise, the nature of the shifts makes a difference with regard to semantic transparency. We assume that some shifts are more costly than others, depending among other things on their general availability as well as their productivity.

Our next step will therefore be to conduct experiments aimed at establishing human ratings for the semantic transparency of AB types, on a continuous scale. We intend to use methodology similar to that used in previous psycholinguistic studies, e.g. Libben et al. (2003), as described in section 3.1.1 of this report. Although Libben et al. (ibid.) used this methodology to create four discrete classes of experimental items, their methodology did in fact create continuous variables in the first instance, from which they extracted their groupings. We will therefore omit the stage of creating groups in order to retain semantic transparency as a continuous variabe. In addition, we will also make at least the following modifications:

1. the experimental items will be randomly selected from a large corpus, and will not therefore be limited to established types

2. all AB types that can fill the frame [det __]$_{\text{NP}}$ will be allowed, since we hypothesise that the distinctions between adjective and noun, and between compound and phrase are themselves gradient

3. some accommodation will be made for the effect of context, possibly by presenting items in their corpus context, or by giving participants the meaning of the combination in that particular context

4. some measure will be taken to ensure that participants are rating the same interpretation of the combination, either by asking them to give a definition, or by providing them with one as described in the previous point.

Once ratings of semantic transparency are established, we will look for distributional correlates, using regression techniques with transparency as the dependent variable and various distributional measurements as predictors. Two main types of distributional measurement will be used: 'morphological' family sizes and co-occurence vectors.

Bell (2012) finds that in English noun-noun combinations, the overall frequency of the AB combination, as well as the morphological family sizes of the constituents, are good predictors of phonological stress placement. However, similarly good predictions can be obtained using a variety of semantic criteria, that on first sight seem unrelated to one another. Bell (ibid.) hypothesises that what unites these semantic criteria is semantic transparency, and that this might also be correlated with the frequency measures, accounting for the similarity of frequency-based and semantic models. We will therefore include measures of frequency and family sizes in our initial models. The notion of 'family size' will be expanded here to include AB combinations that might normally be classed as phrases: the positional family size of A will be the number of B types that follow it within the given frame in a given corpus, and the positional family size of B will be the number of A types that precede it within the given frame in the corpus.

The second type of measurement, co-occurence vectors, capture information about the frequencies with which words occur in the vicinity of other words, within a corpus. According to Harris (1970), differences in meanings between words are correlated with their distribution in the language in question. Indeed, co-occurence vectors have been successfully used within the field of distributional semantics to model various semantic phenomena, including word meaning in context (e.g. Erk & Padó 2008, Thater, Fürstenau & Pinkal 2011) and compositionality (e.g. Guevara 2010, Mitchell & Lapata 2010). Clearly, both the meaning in context of the constituents, A and B, and the degree of compositionality of the combination AB, are relevant to our conception of semantic transparency as summarised in Figure 1, and we will therefore include co-occurence data in our modelling.

Finally, on the assumption that we are able to find distributional correlates of semantic transparency, we will attempt to use these as predictors for stress in English NN combinations and for the availability of AB constituents for anaphora, coordination and modification.

4 Future collaboration with the host institution

As should have become clear from the above, the short visit was only the beginning of the collaboration between the host, Melanie Bell, and the applicant, Martin Schäfer. While continuing our work on the review article, we plan to present the preliminary

results of our work in joint talks at research institutes working on similar and/or related issues. In addition, our intended joint empirical work will necessitate further mutual visits. As mentioned above, we are therefore currently in the process of writing a proposal for the Volkswagen Stiftung.

5 Projected publications/articles resulting or to result from the grant

The review article on semantic transparency will result from this grant.

6 Other comments

We are extremely grateful for the funding through Networds– The European Network on Word Structure. It has really benefitted us, and especially the face-to-face discussions allowed us to clarify a great number of issues. We would therefore welcome any opportunity to contribute and give back to this network in the future.

References

Bell, M. J. (2012), The English NN construct: its prosody and structure, PhD thesis, University of Cambridge.

Erk, K. & Padó, S. (2008), A structured vector space model for word meaning in context, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing', EMNLP '08, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 897–906.
URL: `http://dl.acm.org/citation.cfm?id=1613715.1613831`

Guevara, E. (2010), A regression model of adjective-noun compositionality in distributional semantics, *in* 'Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics', pp. 33–37.

Harris, Z. (1970), Distributional structure, *in* 'Papers in Structural and Transformational Linguistics', pp. 775–794.

Libben, G., Gibson, M., Yoon, Y. B. & Sandra, D. (2003), 'Compound fracture: The role of semantic transparency and morphological headedness', *Brain and Language* **84**, 50–64.

Mitchell, J. & Lapata, M. (2010), 'Composition in distributional models of semantics', *Cognitive Science* **34**(8), 1388–1429.

Thater, S., Fürstenau, H. & Pinkal, M. (2011), Word meaning in context: A simple and effective vector model, *in* 'Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)', Chiang Mai, Thailand.