The variant problem in lexical semantics and translation

Torgrim Solstad
Institut für Maschinelle Sprachverarbeitung, University of Stuttgart
torgrim@ims.uni-stuttgart.de

Lexical semantic analyses necessarily involve more or less explicit considerations concerning the number of interpretational variants of a word form, i.e. identifying the lexical items associated with a lexeme. Likewise, it is mostly of importance to be able to define more precisely the relation(s) between these lexical items, determining whether one is dealing with polysemy, homonymy or other structures of ambiguity. In the following, I will refer to this as the *variant problem in lexical semantics*.

In this paper, I focus on two issues: First, I would like to call *attention to the variant problem*, as it is seldom explicitly addressed in lexical semantics. Establishing the meaning variants of a word form and defining the relation between them is mostly left to the theoretically biased judgement of the individual researcher. Second, and more importantly, I *present a heuristics* which may be considered a first step towards making theory-independent claims about the semantic relations between lexical items. The heuristics makes extensive use of data from parallel corpora.

In many cases, the assumptions about meaning variants are based purely on introspection. This may be unproblematic for classical cases of homonymy such as *bank* ('financial institution' or 'river bank'). It is hardly disputed that these two variants involve separate lexemes. However, in notorious cases of polysemy as often found with prepositions, intuitions are not very helpful, to be witnessed for instance by the dispute over *over* in the framework of Cognitive Grammar (cf. Lakoff 1987, Dewell 1994, Tylor & Evans 2003). What is more, even in cases of thorough empirical investigation the risk of making theoretically biased assumptions is great (c.f. the work of Ruhl 1989 on monosemy). For example, in cognitive-grammatical analyses there is a tendency to always relate as many meaning variants as possible, whereas in model-theoretic semantics there has been a greater readiness to assume homonymy. In conclusion, the variant problem may in general be claimed to be approached in too subjective and theory-dependent ways.

Developing a heruistics for determining the nature of ambiguity of lexical items will not only make for more well-founded lexical semantics analyses, but it will also increase the comparability of such analyses across theoretical frameworks. The heuristics to be presented was introduced by Dyvik (200X) and intended for use in

machine translation and automated reasoning. I will apply it to prepositions, as they have been at the centre of dispute in various approaches to ambiguity in lexical semantics. Since Dyvik and his co-workers have never looked at prepositions, focusing on them also provides a highly relevant test case for the general approach.

Viewing meaning as a relation between two languages, Dyvik argues that using parallel corpora based on original texts and their translations provides for an intersubjectively accessible empirical basis in semantics. After all, translating involves semantic judgements made by highly competent language users in a non-metalinguistic context. I will use German, English and Norwegian data from the Oslo Multilingual Corpus (http://www.hf.uio.no/ilos/OMC/).

Obviously, comparing ambiguities across languages to make statements about the ambiguity of a word form is by itself nothing new in lexical semantics; see e.g. the much quoted paper by Zwicky & Sadock 1975. However, the innovation of Dyvik's approach consists in applying set theory to data from parallel corpora to enable more precise conclusions about meaning relations (parallel corpora have been used in similar tasks in computational linguistics; see Dyvik (200X) for references).

The only epistemological primitive in Dyvik's approach is the *translational relation*: A sign *a* in a source language L1 is paired with a sign *s* in the target language L2 if *a* may be translated by *s* across a reasonable amount of contexts. The translational relation <*a,s*> is assumed to be symmetric and non-transitive. The further application of set theory to the translational relations gives us what Dyvik terms *translational images*. These translational images may in turn be used to make precise characterisations of the ambiguity of a word.

Let me briefly sketch how one proceeds to establish the translational images. For the sake of illustration, I will apply Dyvik's method to the Norwegian preposition *under*, using English as target language. *under* has (at least) spatial and temporal interpretations involving belowness and duration, respectively. The main features of the below procedure are captured in Figure 1 on the next page. Formal definitions will not be included in this abstract. See e.g. Dyvik (200X) for details.

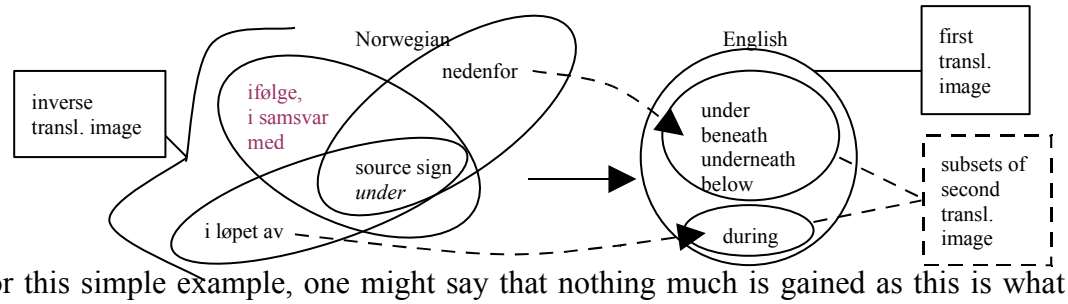
First, all translational relations of Norwegian *under* are extracted from a parallel corpus, be it manually or by way of automatic word-alignment. Signs in a translational relation should belong to the same word class and must occur in comparable contexts. Norwegian *under* enters translational relations with the following English lexemes, reflecting the range of spatial and temporal interpretations: **{under, during, beneath, underneath, below}**. This set is termed the *first translational image* of *under*. The set reflects the possible interpretations of Norwegian *under* and thus gives a first impression of its ambiguity.

Next, all translational relations of the members of the first translational image are extracted, yielding *the inverse translational image* of the source *under*. The result is a set of sets of lexical items in Norwegian (merely hinted at in Figure 1) reflecting the ambiguity of all members of the first translational image, including e.g. *nedenfor*

('below') and *i løpet av* ('during'). In principle, this set may include lexemes not relevant to the interpretation of the Norwegian source sign *under*. For instance, due to an ambiguity of *under* in English, the inverse translational image will include the Norwegian (complex) adpositions *ifølge* and *i samsvar med* with interpretations similar to *according to*. However, *ifølge* and *i samsvar med* do not reflect any ambiguity of Norwegian *under*, but one of English *under*.

In a final step, all translational relations of the lexemes in the inverse translational image of *under* are extracted. The result is termed the *second translational image* of the source sign *under*. The second translational image is usually a rather large set of sets involving many lexemes which are unrelated to the lexeme under investigation. This is due to the non-related ambiguities in the inverse translational image just mentioned. To exclude any ambiguity not related to *under*, the second translational image is restricted to the lexemes already present in the first translational image.

In general, the structure of the restricted second translational image may vary significantly, with different degrees of overlap. The structure may be exploited to define notions like family resemblance and hyperonymy/hyponomy. What is more, it is possible to define non-related ambiguity or homonymy based on it. Simply put, homonymy is defined as the absence of overlap: Taking the union of all sets which have at least one member in common, the following set emerges in the case of Norwegian *under*: **{{under,beneath,underneath,below},{during}}**. The lexemes having spatial belowness interpretations occur in one set, whereas the only lexeme without no spatial interpretation, (temporal) *during*, occurs in a set having no intersections with the first set. Thus, Norwegian *under* is predicted to be two-way ambiguous by this method. Figure 1 summarises the above description:



For this simple example, one might say that nothing much is gained as this is what one would expect. But this would miss the point: What is important is that the ambiguity structure for *under* which corresponds well with intuitions emerged from applying set theory to intersubjectively accessible data.

In addition to the simpler case of Norwegian *under*, I apply the heuristics to the challenging case of German *durch* ('through'), which is highly ambiguous, involving spatial, temporal, agentive, instrumental and causal interpretations.

References

Dewell, R. (1994): *Over* again: Image-schema transformations in semantic analysis. Cognitive Linguistics, 5: 351-380.

Dyvik, H. (200X): Translations as a Semantic Knowledge Source. Ms.

Lakoff, G. (1987): Women, Fire and Dangerous Things. University of Chicago Press.

Ruhl, C. (1989): On Monosemy. State University of New York Press, Albany.

Tyler, A. & V. Evans (2003): The Semantics of English Prepositions. Cambridge University Press, Cambridge.

Zwicky, A. M. & J. M. Sadock (1975): "Ambiguity tests and how to fail them". In J. P. Kimball, ed., Syntax and Semantics, vol. 4, pp. 1-36.