TOWARDS A SEMANTIC MAP OF MOTION VERBS
**Explorative statistical methods applied to a cross-linguistic collection
of contextually-embedded exemplars**

Michael Cysouw (Leipzig, cysouw@eva.mpg.de)
Bernhard Wälchli (Bern, waelchli@isw.unibe.ch)

1. BACKGROUND

SIMILARITY SEMANTICS operates without any reference to the notion of semantic identity (identical meanings, if there is such a thing, are treated as maximally similar meanings). The philosophical basis for a similarity semantics can be found in earlier philosophical literature, especially Mauthner (1923), Berkeley, but also Schopenhauer, Ogden & Richards and K. O. Erdmann. See also Croft (2007).

> "Absolute identity is an abstraction of mathematical thinking. Identity is strong similarity, is a relative notion. The whole conceptualisation in language would not be possible, if we would not be groping in the dark under nothing but fragmentary images and if we would not – because of this fragmentarity – overestimate the similarity and so make a virtue of a vice. The less we know about something, the more we are astounded by similarities […] This is why we use our similarity images or words the more easily the more ignorant we are. Therefore the human language is a consequence of the fact that the human senses are not sharp." (Mauthner 1923: 467, 437-438, translation BW)

Similarity semantics is concerned with the network of similarity and dissimilarity relationships emerging from the cumulative pairwise comparison of concrete situational/contextual settings. Technically, in this framework meaning is a **METRIC ON SITUATIONS**. It does not assume any primitive semantic units. The similarity metric is a non-verbal tool to describe and compare meaning across languages; it is a continuous and empirically obtained alternative to discrete and a priori postulated meta-languages.

2. PRACTICE

In practice, we investigate **CONTEXTUALLY-EMBEDDED SITUATIONS**, which are concrete instantiations (in a particular context) of the functional domain under investigation. We do not consider abstract functional domains, but only concrete examples of such a domain. An expression of a situation in a particular language is a **CONTEXTUALLY-SITUATED EXEMPLAR**.

Similarity semantics is based on the cross-linguistic interpretation of the **ICONICITY PRINCIPLE** "recurrent identity of form between different grammatical categories will always reflect some perceived similarity in communicative function." (Haiman 1985: 19). If two situations are recurrently expressed using the same formal expression, in language after language, than these two situations are considered similar.

A **LANGUAGE-PARTICULAR FORM-CLASS** ("a class of situations using the same form"), cf. "language-particular category" (Haspelmath 2007), is a set of contextually-embedded situations in which a language-particular form (lexeme, morpheme, or construction) can be used. We consider this set of possible uses to be the "meaning" of the language-particular form.

3. DATA

In similarity semantics **THERE IS NO ABSOLUTE IDENTITY**. However, the practical application of the theory acts as if cross-linguistic translation equivalents were identical when they are in fact only very similar. This works to the extent that the cross-linguistic semantic differences between the identified exemplars are smaller than the semantic difference between the sampled situations within a language. This requires that the situations are defined very sharply which is ascertained by their contextual embeddedness.

MASSIVELY PARALLEL TEXTS (Cysouw & Wälchli 2007) allow for a direct cross-linguistic comparison of contextually embedded situations without previous abstraction of language-particular systems and without previous classification of semantic contexts. This makes it possible to compile large databases of cross-

linguistically comparable examples in a large number of diverse languages—at the cost of restricted idiomaticity due to translation. The abstract idea of translational equivalence, pervasive in functional linguistics, is implemented in practice with all practical problems of translation implied.

Another, more costly, method of collecting cross-linguistic examples of contextually-embedded situations is the use of **TRANSLATIONAL QUESTIONNAIRES** (Dahl 1985) or **VISUAL QUESTIONNAIRES** (Levinson & Meira 2003). These approaches emphasise the importance of high quality data elicited under experimental conditions in the natural environment of the speakers of a language.

The entities to be compared are particular texts in (or about) a particular language rather than an abstract language in general. We refer to this conception of 'language' as **DOCULECT** ("documented lect").

4. ANALYSIS

The data in the resulting large database is too diverse to allow for a manual extraction of generalisations. General tendencies are extracted by means of **EXPLORATIVE STATISTICAL METHODS** (clustering, dimensional scaling). The same procedures are always applied to all data in the database. There are no "exceptions" separated *a priori*.

There are (at least) **THREE LEVELS OF ANALYSIS**. We investigate the relation between:
 – contextually-embedded situations ("situations", "examples")
 – language-specific form-classes ("lexicalisations", "categories")
 – doculects ("texts", "languages")

For all three levels, a common procedure is to calculate first a **[DIS]SIMILARITY MATRIX** and then to analyse and/or visualise the structure of this matrix. It is important to realise that there is never a 1:1 relationship between the data and the visualisation: there is always some information lost (and added) in the process.

5. REFERENCES

Croft, William, & Poole, Keith T. 2006. Inferring universals from grammatical variation: multidimensional scaling for typological analysis. To appear in *Theoretical Linguistics.*

Croft, William. 2007. Exemplar Semantics. Draft. http://www.unm.edu/~wcroft/Papers/CSDL8-paper.pdf

Cysouw, Michael. 2007. Building semantic maps: the case of person marking. In M. Miestamo, & B. Wälchli (eds.), *New Challenges in Typology: Broadening the Horizons and Redefining the Foundations*. Berlin: de Gruyter.

Cysouw, Michael, & Bernhard Wälchli (eds.). 2007. Parallel Texts. Using translational equivalents in linguistic typology. Theme issue in *Sprachtypologie & Universalienforschung*. http://email.eva.mpg.de/~cysouw/pdf/STUF.pdf

Dahl, Östen. 1985. *Tense and Aspect Systems*. Oxford: Blackwell.

Haiman, John. 1985. *Natural syntax*. Cambridge: Cambridge University Press.

Haspelmath, Martin. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology* 11(1): 119-132.

Levinson, Stephen & Meira, Sérgio. 2003. 'Natural concepts' in the spatial topological domain-adpositional meanings in crosslinguistic perspective: an exercise in semantic typology. *Language* 79: 485-516.

Mauthner, Fritz. 1923 [1982]. *Beiträge zu einer Kritik der Sprache.* Erster Band: Zur Sprache und zur Psychologie. 2., vermehrte Auflage. Leipzig: Meiner / Frankfurt: Ullstein Materialien.

Wälchli, Bernhard. 2007. Constructing semantic maps from parallel text data. Draft. http://ling.uni-konstanz.de/pages/home/a20_11/waelchli/waelchli-semmaps.pdf.