Thomas Meyer (Idiap Research Institute, Martigny, Switzerland), Charlotte Roze (Alpage Group, INRIA and University of Paris VII, France), Bruno Cartoni (Department of Linguistics, University of Geneva, Switzerland), Laurence Danlos (Alpage Group, INRIA and University of Paris VII, France), Sandrine Zufferey (Department of Linguistics, University of Geneva, Switzerland), and Andrei Popescu-Belis (Idiap Research Institute, Martigny, Switzerland)

Disambiguating discourse connectives using parallel corpora: senses vs. translations

Discourse connectives are words or phrases that indicate senses holding between two spans of text. The theoretical approaches accounting for these senses, such as text coherence, cohesion, or rhetorical structure theory, share at least one common feature: they acknowledge that many connectives can indicate different senses depending on their context. For instance, in English, ‘while’ can sometimes indicate a temporal sense, but other times a comparison, an opposition, or a concession. Depending on its sense, the translation of a connective into another language can vary greatly, either using an equivalent connective, or using a different construction or even no explicit connective at all.

The objective of this study is to characterize the multifunctionality of a subset of connectives which are both, frequent and ambiguous. We will define the main senses of each connective, describe a reference annotation of connectives with their senses in parallel corpora, and make quantitative observations on the frequencies of senses and their translations. The parallel texts are English/French parliamentary debates (with known source language and its direct translation) from the Europarl (Koehn, 2005) and Hansard (Roukos et al., 1995) corpora.

Two possible approaches to corpus-based studies of connectives have been explored in the past. Our objective is to show that combining the two produces richer and more reliable results.

The first approach provides annotators with descriptions of the possible senses of each connective, and requires them to label each occurrence with one sense, as in the English Penn Discourse Treebank (Prasad et. al., 2008). Similarly, Roze et. al. (2010) have identified possible senses of French connectives in the LexConn database, with 328 connectives totaling 428 possible senses. The senses and their definitions are currently used for annotating English and French texts.

The second approach considers the translations of connectives observed in parallel corpora – e.g. like in the study of causal connectives in French/Dutch novels by Denturck (2010). Our observations on temporal/contrastive or causal connectives show that beyond the large variety of possible translations, there are dominant clusters of translations corresponding to the main senses identified monolingually. For instance, the French connective ‘alors que’ has four frequent translations into English in the Hansard corpus: ca. 50% ‘even though’, 10% ‘when’, 5% ‘given that’, and 10% of no direct lexical equivalent. These translations reflect its multifunctionality as an indicator of concession or a temporal sense. We will present findings for temporal/contrastive and causal connectives such as ‘while’, ‘since’ in English and ‘alors que’, ‘en effet’, 'parce que', 'car', and 'puisque' in French, with respect to use in original texts and their translations.

As a result, a multilingual database of connectives will be constructed, including descriptions of their senses and principal translations, augmented with frequency information from parallel corpora. The annotated resource will be used for training and testing an automatic system that disambiguates connectives, as a preliminary stage to their automatic translation.

References:

Denturck, Kathelijne (2010): Translation universals: the case of causal connectives in French and Dutch translations. A corpus-based study. Workshop Connectives across Languages: Explicitation and Grammaticalization of Contigency Relations. http://www.francais.ugent.be/index.php?id=25&type=file [16.11.2010].

Koehn, Philipp (2005): Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit 2005.

Roze, Charlotte, Laurence Danlos and Phillippe Muller (2010): LEXCONN: a French Lexicon of Discourse Connectives. Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010).

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber (2008): The Penn Discourse Treebank 2.0. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Roukos, Salim, David Graff, and Dan Melamed (1995): Hansard French/English. Linguistic Data Consortium, Philadelphia.