Abs-188

Lydia-Mai Ho-Dac, Cécile Fabre, Marie-Paule Péry-Woodley, Josette Rebeyrolle, and Ludovic Tanguy
CLLE-ERSS (CNRS & Université de Toulouse)

# High-level discourse structures: topical chains and enumerative structures in a diversified annotated corpus

One of the outcomes of the ANNODIS project (Ho-Dac et al 2009, 2010) is a diversified corpus annotated with two frequent textual motifs: topical chains – TCs – and enumerative structures – ESs. The corpus has been manually annotated with both the motifs and the clues signalling them. These data can now be exploited in a comparative mode in order to examine TCs and ESs in the three sub-corpora: 1) reports in the field of international relations; 2) scientific articles (proceedings of a linguistics conference); 3) encyclopaedia articles (from Wikipedia).

The initial step is to take a quantitative look at each motif: composition, distribution, and match with document structure (Power et al 2003). Though the motifs are common in all three corpora, differences appear in their frequency, in their length and coverage (proportion of text involved), in their composition (for ESs: number of items, presence of a trigger and/or closure). Another important aspect is their granularity: this notion is approximated via a typology in which types correspond to different forms of interaction between the motifs and the document's layout structure (sections and headings, formatted lists and paragraphs).

We then examine the data from several qualitative angles in order to arrive at a functional characterisation of the motifs. Of special interest to us is the link between particular forms of signalling and specific functions: ESs with items introduced by sequencers, for instance, are functionally different from ESs whose items are introduced by circumstantial adverbials. A continuum is proposed from ESs signalled by purely textual cues (e.g. bullet points) to ESs whose cues carry ideational contents (such as adverbials) (Halliday 1977). The different corpora are compared in terms of the functional classes and their linguistic correlates, as illustrated by the table below:

| **Corpus** | **Number of ESs** | **ESs/text** | **Coverage** | **Granularity** | | | |
|---|---|---|---|---|---|---|---|
| | | | **(% of text)** | **sections** | **formatted lists** | **multi-paragraph** | **intra-paragraph** |
| WIKI | 332 | 13.28 | 53.77% | 19,28% | 39,16% | 20,78% | 20,78% |
| CMLF | 263 | 11.95 | 47.91% | 9,13% | 23,19% | 26,62% | 41,06% |
| GEOPO | 234 | 9 | 28.73% | 6,84% | 10,26% | 20,94% | 61,97% |
| mean | | 11.36 | 43.09% | 12,55% | 25,93% | 22,68% | 38,84% |

Table 1: Frequency and inter-corpus variations for enumerative structures (ESs)

Finally, the two motifs are observed in context and in their interaction. A special case of interaction concerns ESs interacting with themselves via recursivity, a remarkably frequent occurrence in our corpus. This analysis of how the motifs behave in text also leads to cross-corpus comparisons.

**References**

Halliday, M.A.K. (1977/2003) ‘Text as semantic choice in social contexts’. In *The Collected Works of*

*M.A.K. Halliday (Volume 2): Linguistic Studies of Text and Discourse*, Jonathan Webster (ed.), 23–81. London: Continuum.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P., Rebeyrolle, J. & Tanguy, L. (2009) 'A top-down approach to discourse-level annotation', *Corpus Linguistics Conference 2009*, 20-23 July, Liverpool, UK.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P. & Rebeyrolle, J. (2010). 'On the signalling of multi-level discourse structures', *MAD 2010 : Multidisciplinary Perspectives on Signalling Text Organisation*, Moissac (France) 17-20 mars 2010, 94-105.

Power, R., Scott, D. & Bouayad-Agha, N. (2003). 'Document Structure'. *Computational Linguistics* 29(2), 211-260.