The annotation of discourse markers in an English-Hungarian parallel corpus: parsing out the complex relationship between formal and functional features

Peter Furko
Károli Gáspár University of the Reformed Church in Hungary
4 Reviczky Street, Budapest, Hungary, H-1088
Email: furko.peter@gmail.com

We can find the most striking differences between written and spoken discourse-relational devices (DRDs) in terms of their respective multifunctionality, context-dependence, implicitness and type-token ratios. As a result, the most challenging task for discourse annotators is to tag a set of highly frequent DRDs such as well, you know, I mean, I think etc., which are used in a wide range of contexts with numerous (discourse-relational and interpersonal) functions, rather than DRDs such as in other words, or rather, in short, etc. which mark more explicit relations between discourse segments and are used with higher type/token ratios.

The present paper will focus on of course and its Hungarian counterparts persze (~'sure'), természetesen (~'naturally'), naná (~inf./stigm. 'sure'), and will argue that of course, similarly to well, you know, I mean, etc. is a discourse marker (DM) that has no fixed meaning but a meaning potential compatible with a variety of core (basic) and peripheral functions (cf. Bell, 1998).

There is general agreement in the literature that a contrastive analysis can help tease out the diversity of meaning relations that semantically bleached DMs mark (cf. Mortier & Degand, 2009), while there is also an increasing awareness of the “indexically rich” situational meaning of DMs (Aijmer, 2013) and the resulting need to analyse DMs across a variety of speech situations and genres. In order to meet both demands, the study will map the functional spectrum of of course in private and public conversations (based on previous research) as well as in a one-million-word English-Hungarian parallel corpus which is based on dramatized dialogues.

The source language (English) DMs have been aligned with the target language (Hungarian) lexical items (alternatively, the absence of a translation equivalent has been noted), while individual tokens of of course have been annotated for the following formal properties:

- pausing, prosody or accent;
- DM clusters, collocations, lexical co-occurrence patterns;
- speaker roles (interviewer, interviewee, equal/unequal encounters);
- speech act of the host utterance / preceding utterance;
- position in the utterance (initial, medial, final);
- position in the turn (initial, medial, final);
- the host unit’s position in conversational structure (first or second part of an adjacency pair, embedded sequence, etc.).

In terms of functional properties, the following features will be annotated:

- primary (discourse-relational) functions (addition, concession, consequence, or other);
- secondary (interpersonal and interactional) functions (feedback, stance, +/- politeness, hedging, boosting);
- the annotator’s confidence in identifying discourse relations marked by of course (on a scale of 1-3);
- the annotator’s confidence in identifying interpersonal and interactional functions fulfilled by of course (on a scale of 1-3).

The paper will focus on answering the following research questions:

RQ1: Is there a correlation between the (primary/secondary) function of source language DMs and target language items?
RQ2: Is there a correlation between an explicit (primary or secondary) function and one or more of the formal features?

The preliminary results, based on two types of ELAN corpus queries (find overlapping labels, N-gram with annotations), suggest that DM clusters, lexical collocations and target language items are the most reliable formal indicators of the textual and interpersonal functions of course marks, while prosodic and positional features, genre, speaker roles and the SA of the host unit and/or adjacent DU might also increase the likelihood of of course realizing particular discourse relational or interpersonal/interactional functions.
References

