

Issues of segmentation for the cross-linguistic study of discourse connectives

Maité DUPONT¹

Université catholique de Louvain¹

In research on discourse relations, it is generally accepted that the minimal discourse unit corresponds to the clause. This means that – at least in written language – a given connective is considered to perform a discourse function when it links two clauses (or larger units), while a linking relation expressed below clause level (as in (1) *Paul and Mary are going to the beach*) is not viewed as a discourse relation (see e.g. Miltsakaki et al. 2004; Mann & Thompson 1987). However, the exact nature and delimitation of the clause itself is little discussed in the literature on text structuring (Muller et al. 2012) and the criterion used to grant clause status to a given text segment is often limited to whether or not it contains a verb phrase. This presupposes that the definition of the clause is straightforward, which is far from being the case, especially when considered in a contrastive perspective. For the English-French language pair, for example, the definitions of the clause differ markedly in the literature: while in English the definition of the clause is very much centred on the notion of verb, with the clause defined as “any stretch of language centred around a verbal group” (Thompson, 2014: 17), in French the definition of the clause is much less reliant on the notion of verb, and the clause is said to be centred around a predicate, which may be verbless (see e.g. Le Goffic 1994; Lefeuve 1999). Therefore, a study investigating discourse connectives in English and French that would only include connectives linking two verbal units would run the risk of overlooking some relevant uses of French connectives, such as in (2) *Luc a dû démissionner. Une véritable tragédie donc pour lui* (Danlos et al. 2015), where *donc* qualifies as a discourse connective, as its host segment is viewed as a clause in its own right. In addition, even English displays some verbless structures with clausal status (such as in (3) *Although always helpful, he was not much liked*, Quirk et al. 1985: 992), and a definition of the clause fully centred on the verb overlooks such relevant structures.

This paper sets out to explore the issue of discourse segmentation for the cross-linguistic study of discourse relations in written language, focusing on connectives of contrast in English and in French (e.g. *however, whereas, but; cependant, alors que, mais*). The study is based on a comparable corpus of quality newspaper editorials (c. 2 million words per language, compiled from quality newspapers including *The Guardian, The Observer; Le Monde, Le Figaro*, etc.). In order to establish clearly what items can and cannot be considered as discourse connectives in the corpus, I have established (i) a broad classification of the types of clauses which holds for both English and French, thus allowing for the inclusion of some verbless structures as genuine discourse units, and (ii) a set of four criteria to distinguish between true verbless discourse units and non-discourse units at phrase level, a distinction which is particularly difficult to make, for example in the case of verbal ellipsis. The criteria rely on questions such as the following: ‘do the two segments linked perform the same grammatical function in the sentence’; or ‘is there a strong punctuation mark between the two units related?’ Based on these four criteria, I put forward a three-step decision process to determine with as much certainty as possible whether or not a given item can be considered as a discourse connective with respect to the nature of the segments that it links. The study also discusses a set of remaining borderline cases, which could not be handled even with the application of the aforementioned criteria, and

for which a “phrasal-clausal” category had to be created. This phrasal-clausal category is a witness to the great variety and complexity of the syntactic structures in which connectives can be involved, which is not fully grasped when using purely verb-based criteria of selection.

In a second stage of the study, I have operationalized my selection criteria by annotating the connectives in my corpus on the basis of the classification of clause types designed for the selection of connective items. The objective was to determine whether English and French display differences in the types of syntactic structures in which they tend to use connectives. The coding system designed for that purpose is made up of three layers, viz. (i) part-of-speech of the connective (i.e. adverb, coordinator, subordinator); (ii) syntactic type of host clause (i.e. finite, non-finite or verbless) and (iii) rank status of the host clause (i.e. main, hypotactic or embedded). The results show that English and French display striking differences in connective use. For instance, while English tends to use subordinators in verbless hypotactic clauses (e.g. ***Even if** occasionally right, this is hard to take*), in French this pattern is quite infrequent. Instead, the French data displays a significantly more frequent use of adverbs and coordinators in verbless clauses than English (e.g. *Pas un mot **en revanche** du tibérisme comme système/ La réforme est enfin sur le point d'être engagée. **Mais** sur de très mauvaises bases*). These cross-linguistic differences provide further justification for the inclusion of verbless discourse segments in the analysis.

References

- Danlos, L., M. Colinet and J. Steinlin. (2015): FDTB1: Annotation des connecteurs de discours dans un corpus français. Manuel d'annotation du FDTB. Available at https://gforge.inria.fr/frs/?group_id=6145 [last accessed 15/03/2017]
- Lefeuve, F. (1999): *La phrase averbale en français*. Paris: L'Harmattan.
- Le Goffic, P. (1994): *Grammaire de la phrase française*. Paris: Hachette Éducation.
- Mann, W. and S. Thompson. (1987): Rhetorical Structure Theory: a theory of text organization. USC Information Sciences Institute, Technical Report ISI/RS-87-190.
- Miltsakaki, R., R. Prasad, A. Joshi and B. Webber. (2004): The Penn Discourse Treebank. In *Proceedings of the 4th Language Resources and Evaluation Conference*, Lisbon, Portugal.
- Muller, P., Vergez-Couret, M., Prévot, L., Asher, N., Farah, B., Bras, M., Le Draoulec, A. and Vieu, L. (2012): Manuel d'annotation en relations de discours du projet ANNODIS. Available at <http://redac.univ-tlse2.fr/corpus/annodis/> [last accessed 15/03/2017].
- Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985): *A Comprehensive Grammar of the English Language*. London: Longman.
- Thompson, G. (2014): *Introducing Functional Grammar*. London and New York: Routledge.