Connectives in comparative perspective: Multilingual corpora and crosslinguistic databases

1 Connectives as a challenge for linguistic theory and application

Connectives such as therefore, because, due to, etc. are functional elements that connect text segments of different sizes, e.g. (sets of) sentences or clauses. Like markers of coreference, they constitute an important device for signalling coherence in a text. In particular, they establish semantic or pragmatic relations between the elements in question, e.g. a relation of causation, condition, contrast, elaboration, etc. (cf. Hasan and Halliday 1976; Sanders 1997; Sæbø 2011).

From a syntactic point of view, two major types of sentence connectives are traditionally distinguished: coordinating ones (e.g. and, but; cf. Haspelmath 2007) and subordinating ones (e.g. because, although; cf. Kortmann 1997 on adverbial subordinators in European languages, and several contributions to Fabricius-Hansen and Ramm 2008 for discussion of the subordination-coordination dichotomy or continuum). On the discourse level, certain adverb(ial)s also often have a connecting function, e.g. the so-called ‘conjunctions’ of English (Quirk et al. 1985, 631ff, e.g. nonetheless) and the ‘Konjunktionaladverbien’ of German (e.g. allerdings, jedoch; cf. Brauše 1983, among others). Finally, there is a debate whether or not certain prepositions (e.g. trotz, wegen / despite, due to) should also be regarded as forming a natural class with sentence connectives and discourse connectives like those mentioned above. In our view, the inclusion of prepositions in the class of connectives is justified when they take propositional arguments, not least because of the widespread polyfunctionality that can be observed in this domain, as in the case of since, for instance, which functions both as a preposition and as an adverbal subordinator.
The class of connectives, thus defined as a syntactically heterogeneous group, presents a series of interesting and non-trivial challenges for both linguistic theory and areas of applied linguistics such as translation and language teaching. The relationship between connectives and meanings is many-to-many, i.e. a specific discourse relation such as ‘concession’ can be expressed with several connectives (e.g. German obwohl, dennoch, trotzdem, nichtsdestoweniger, trotz, among others; cf. Breindl 2004), and specific connectives often have several interpretations (e.g., schließlich has a temporal and an argumentative reading; cf. Weydt 1979 for a detailed study of the latter function). The elements of this class tend to be polysemous or vague (see for instance Kortmann 1997; Blühdorn 2010 on the polyfunctionality and a semantic typology of adverbial subordinators), and their interpretation often varies with the context. This high degree of context dependence leads to considerable difficulties for both lexicography and translation – in particular, machine translation – as the selection of a specific translation for a given connective makes reference to contextual information and often requires pragmatic inferencing. Moreover, connectives tend to be highly register-sensitive (cf. Zufferey 2011). In view of such complications, it comes as no surprise that connectives also present enormous difficulties for both first and second language learners (see e.g. Cain and Nash 2011).

Given the many-to-many relationship between connectives and the semantic relations that they express, dictionaries and grammars do not provide adequate formats for the representation of their semantic properties. A dictionary will normally provide a list of possible translations for a given item, i.e. it will typically be organized semasiologically. A grammar may be structured onomasiologically and provide the linguistic means available for any given meaning (cf. the presentation of adverbial clauses in Zifonun et al. 1997). The very format of a print publication restricts the possibilities of representing form-to-function mappings, however, and the many-to-many relationship between connectives and discourse relations cannot easily be reproduced in them.

As an alternative to printed reference works, several attempts have been made to represent the syntactic and semantic properties of connectives in different electronic formats, e.g. in relational databases or electronic lexicons like DiMLex (Stede 2002) and LEXCONN (Roze et al. 2010). Databases allow for a flexible retrieval of (multidimensional) information and can moreover be used for the storage of significantly larger amounts of data than print publications. Moreover, they are dynamic and can thus accommodate the growing pool of linguistic resources available, in particular corpus data. Ideally, the relevant databases should access corpora directly and thus allow for the retrieval of various types of quantitative information according
to the requirements of any given user. Finally, given the context-sensitivity of connectives, generalizations concerning their semantic and structural properties will often be probabilistic, and such probabilistic information can be provided more easily in dynamic databases than in books. It is this research programme that provides a frame of reference for the envisaged workshop on ‘Connectives in comparative perspective: Multilingual corpora and cross-linguistic databases’.

2 The multilingual study of connectives: Challenges and desiderata

In recent years, considerable progress has been made in the development of monolingual resources for the description and study of connectives. For German, a major milestone was published in 2003, i.e. the *Handbuch der deutschen Konnektoren* (Vol. 1, Pasch et al. 2003), which focuses on the syntax of connectives. This volume will soon be followed by a second one on semantics. Furthermore, some electronic resources like DiMLex and LEXCONN have become available, as well as an English corpus that is richly annotated for connectives and their relations, the Penn Discourse Tree Bank/PDTB (cf. Prasad et al. 2008). Several PDTB-inspired follow-up projects for other languages (among them Arabic and Turkish) are currently under way. Furthermore, there are a few projects aiming at a systematic correlation between the annotation of sentence syntax and that of discourse structure, notably in the Prague Dependency Treebank\(^1\) and in the Copenhagen Dependency Treebank.\(^2\)

However, a comparable infrastructure for detailed and thorough *multilingual* research on connectives is not yet available. The development of such an infrastructure – no doubt a major desideratum of both research on connectives and specific linguistic applications like translation and foreign-language teaching – is a challenging undertaking. First of all, the multifactorial study of connectives and their interpretations in context requires suitable cross-linguistic corpora. Even though comparative corpus linguistics has made some progress in recent years, and a number of parallel and comparable corpora are available (cf. Aijmer 2008), the amount of data and, even more importantly, the degree of granularity in corpus annotation needed for robust comparative generalizations goes far beyond the corpus resources available at present. What is needed, thus, is a multilingual corpus that is appropriate for the

\(^1\)
http://ufal.mff.cuni.cz/pdt2.0/

\(^2\)
http://code.google.com/p/copenhagen-dependency-treebank/wiki/CDT
detailed investigation of connectives; and this programme requires richer annotations than are available, for instance, in the PDTB.

Given this goal, the multilingual study of connectives involves at least the following prerequisites:

- A multilingual corpus needs to be compiled, which is to some extent balanced in terms of register/genre, thus allowing for crosslinguistic comparison; this corpus should include both parallel texts and comparable texts.

- The multilingual corpus needs to be annotated. The development of a system of variables (tag set, ontology) for the annotation of multilingual corpora is, in itself, a difficult task. The first step in the development of an infrastructure as described above thus consists in the development of an ontology of intersentential or discourse relations, as well as a system of other variables required for the analysis of connectives (properties of the clausal arguments, such as tense/aspect, modality, etc.).

- The required annotation can partly be done automatically, but specific (especially semantic and pragmatic) variables have to be annotated manually. Given that annotations at several levels are required, appropriate software for multi-level annotation is needed.

Once a richly annotated multilingual corpus is available, relational databases can be generated which contain information about the complex relationships between the structural properties of connectives and their interpretation in specific types of context. Alternatively, a relational database is not even necessary if a sufficiently powerful and user-friendly search interface for the corpus is provided. Rather than looking up a connective in a dictionary or database, users will then be able to search for instances of connectives with specific interpretations and in specific types of context, as well as their translations into other languages in comparable types of context. Given that such a direct-access approach is a rather ambitious goal whose implementation will require collaborative long-term projects, the more ‘standard’ approach based on relational databases with information on connectives should not be dispensed with for the time being.

3 Objectives of the workshop

The corpus-based crosslinguistic study of connectives as outlined in Section 2 is a major challenge that can only be met by an international team of
The envisaged workshop intends to bring together such experts and to form a consortium which will pursue this programme over the next few years. First steps towards the establishment of such a group have been made already. A workshop on multilingual databases of connectives (funded by the SNF) was organized by B. Cartoni and S. Zufferey (Geneva) in Les Diablerets in autumn 2011. We intend to continue the research strategy devised at this workshop, giving it a slightly different direction by focusing on corpora rather than databases, and by inviting colleagues who have not so far been involved in this international enterprise.

The workshop will focus on four topic areas:

1. **Theory**: the development of an annotation scheme that is applicable to all languages under consideration (and ideally universally valid), including an ontology of discourse relations and a system of variables required for studying the properties of connectives, arguments, and broader context.

2. **Methodology**: the compilation of multilingual corpora that is adequate for the comparative study of connectives, enriched with meta-data relevant to the topics of inquiry, as well as statistical methods for the exploration of such corpora.

3. **Infrastructure**: the development of software for the annotation and exploration of multilingual corpora.

4. **Application**: the use of corpora for specific problems of translation (especially machine translation).

**Literature**


