In contrast – A Complex Discourse Connective

Erhard W. Hinrichs and Monica L. Lău
Universität Tübingen
{eh,mlau}@sfs.uni-tuebingen.de

1 Introduction

The semantics and pragmatics of discourse structure has been a central theme in linguistic research for quite some time. Recent research on large-scale annotation of discourse relations for the purposes of natural language processing applications has resulted in new insights in the properties of such relations and in concrete proposals on how to annotate them. A particularly ambitious and interesting effort of this kind is the Penn Discourse Treebank (PDTB), a corpus of 1 million words which is being annotated for discourse connectives and their arguments, more specifically for connectives such as *but*, *because*, *after*, and *when* that are either realized lexically (explicit connectives) or that have no overt linguistic realization, but that can be inferred as a logical relation between pieces of discourse (implicit connectives).

The detailed PDTB annotations, which by now comprise a substantial corpus of linguistic data, make it possible to revisit an open research question that had been raised repeatedly in the literature, albeit without yielding concrete results. This open research question concerns the similarities and differences between syntactic and semantic relations at the sentence level and at the discourse level. Webber (2006) and Lee et al. (2006) have addressed this very issue in the context of the PDTB annotations and have arrived at the following empirical generalizations:

(1) While the arity of predicates at the sentential level can vary, e.g. one argument in the case of intransitive verbs, two in the case of transitives, three for ditransitives, etc., the arity of discourse connectives is fixed and consists of exactly two arguments.

(2) While syntactic dependencies can be quite complex and may involve highly nested or even crossing dependencies of various kinds, dependencies expressed by discourse connectives tend to be much more limited, typically involving tree-like structures and not introducing structural ambiguities of scope or attachment.

(3) More complex cases of discourse connectives that prima facie seem to involve crossing or partially overlapping arguments can be reduced to independent discourse mechanisms of anaphora and attribution and thus do not introduce any added complexities.

The purpose of this paper is to further examine and refine the above hypotheses by looking in some detail at a family of discourse connectives, all involving the notion of contrast. It turns out that this family of connectives cannot easily be accommodated by the above generalizations because it displays a number of properties characteristic of sentential dependencies, but purportedly not expressible by discourse connectives.

2 The Data

The British National Corpus (BNC) served as the data source for the present investigation. The BNC is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. The reasons for choosing the BNC rather than the Wall Street Journal (WSJ) corpus, which provides the data source for the PDTB, are two-fold: (i) The BNC is a hundred times larger than the 1-million word WSJ corpus and thus yields a much larger data source, and (ii) the BNC is much more balanced in the genres represented than the WSJ. The lemma *contrast* with part of speech tag *noun* appears 6816 times in the BNC. In the current experiment we extracted all occurrences of the noun sense of *contrast* in combination with the preposition *in* and possibly intervening adjectives such as *profound*, *sharp* or *stark*, yielding patterns such as *in contrast* or *in sharp contrast*. While additional data involving the preposition *by* or related connectives such as *in comparison* or *by comparison* still need to be examined, the current data set of 2492 examples of the phrase *in (ADJ) contrast* suffices to address the theoretical issues most relevant for this paper.

3 Empirical Findings and Theoretical Conclusions

Our empirical findings will be based only on clause-initial cases of *in* [*sharp*, *stark*, *marked*, ...] *contrast*, but currently do not cover occurrences in copula constructions, such as [*is*, *seems*] *in contrast with*. One of the properties that makes the connective *in contrast* worthy of special consideration is the fact that it has both intra-sentential and extra-sentential uses. The former arises when the connective combines with the preposition *to*.

(1) [In contrast] [$_{arg1}$ to his predecessors who worked at all hours of the day]

[$_{arg2}$ Macmillan tended to keep office hours] . B0H(0476)

In such prepositional usages, the argument structure seems to be the same as for other simple discourse connective in the sense that it has exactly two distinct and non-overlapping arguments, shown as *arg1* and *arg2* in (1). However, upon closer inspection, there is a notion of parallelism between subparts of each argument that a more fine-grained annotation should reflect. As shown in (1), this parallelism interestingly involves crossing dependencies. In Lee et al.'s and Webber's view, the existence of such crossing dependencies is characteristic of (intra-)sentential connectives. However in the case of the *in contrast* connective, they also carry over to discourse uses as in (2).

(2) It's a shame, then, that [$_{arg1}$ its gearchange is coarse and sloppy. [In contrast],

[$_{arg2}$ the Calibra's is light and quick], although the clutch action could be more progressive]. A6W(0763)

The same type of crossing parallelism exhibited in (1) now involves material across clauses. While (2) involves material in adjacent clauses, there are plenty of examples where such dependencies extend over an entire paragraph or over even larger amounts of text. While Lee et al. and Webber recognize the existence of such long distance dependencies among the arguments of discourse connectives, they invoke the notion of *discourse anaphora* (in the sense of Hinrichs 1986 and of Webber et al. 2003) as a more general mechanism to account for such cases. However, the role parallelism exhibited by the connective *in contrast* cannot be subsumed under discourse anaphora for the following reasons: (i) Discourse anaphora depends solely on the notion of *discourse salience* and is thus not structural in nature. In the case of the contrast relation, however, it is precisely this parallelism of structure that is required. In this regard, the discourse connective *in contrast* behaves rather like the syntactic construction of gapping. (ii) Discourse anaphora typically refers to a relation between exactly one anaphor and its antecedent. However, *in contrast* crucially involves at least two contrast pairs.

The more fine-grained semantic representation of the discourse connective *in contrast*, as shown graphically in (2) can be symbolized as in (3):

(3) in_contrast(X's gearchange has property Y, <Corrado, coarse and sloppy>,<Calibra, light and quick>)

(3) shows a three-argument relation between a comparison pattern, which contains free variables for the contrast pairs rendered as tuples in argument positions 2 and 3. As such, the representation in (3) constitutes a counterexample to Lee et al.'s assumption that discourse relations are always of arity 2.

4 Conclusion and Future Work

In sum, the discourse connective *in contrast* cannot be easily accounted for by the empirical generalizations about discourse relations put forth by Lee et al. (2006) and by Webber (2006). In fact, this connective violates all three empirical generalizations which have been stated by these authors. Thus, the properties exhibited by this connective show that, at least in the limiting case, the argument structure of discourse connectives can be just as complex as genuine syntactic dependencies.

In future work we plan to consider a wider range of contrast relations in discourse in order to ascertain whether the properties of the discourse connective *in contrast* will generalize to these cases as well. A second line of research will investigate ways of automatically detecting comparison patterns and contrast pairs, which are exemplified in (3) above, by means of machine learning techniques.

References

Burnage, A.G. and G. Baguley (1996). The British National Corpus. Library and Information Briefings 65.

Hinrichs, E. (1986). Temporal Anaphora in Discourses of English. Linguistics and Philosophy 9(1): 63-82.

Lee, A., R. Prasad, A. Joshi, N. Dinesh and B. Webber (2006). Complexity of Dependencies in Discourse: Are Dependencies in Discourse More Complex Than in Syntax? Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories. Prague, Czech Republic.

Webber, B. (2006). Accounting for Discourse Relations: Constituency and Dependency. In M. Butt, M. Dalrymple, and T. King, Intelligent Linguistic Architectures, Stanford: CSLI Publications, 2006, pp. 339-360.

Webber, B., A. Joshi, M. Stone, and A. Knott (2003). Anaphora and Discourse Structure. Computational Linguistics 29(4): 545-587.