In contrast - A Complex Discourse Connective

Erhard Hinrichs and Monica L. Lău

Seminar für Sprachwissenschaft/Computerlinguistik Universität Tübingen {eh,mlau}@sfs.uni-tuebingen.de

Abstract

This paper presents a corpus-based study of the discourse connective *in contrast*. The corpus data are drawn from the British National Corpus (BNC) and are analyzed at the levels of syntax, discourse structure, and compositional semantics. Following Webber et al. (2003), the paper argues that *in contrast* crucially involves discourse anaphora and, thus, resembles other discourse adverbials such as *then*, *otherwise*, and *nevertheless*. The compositional semantics proposed for other discourse connectives, however, does not straightforwardly generalize to *in contrast*, for which the notions of contrast pairs and contrast properties are essential.

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*, *for example*, *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).

On the basis of the detailed PDTB annotations, which by now comprise a substantial corpus of linguistic data, it has become 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:

- 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 the independent discourse mechanisms of anaphora and attribution and thus do not introduce any added complexities.

The third generalization is further elaborated by Webber et al. (2003) who distinguish between coordinating conjunctions such as *and*, *or*, *so*, and *but* and subordinating conjunctions such as *although*, *whereas*, and *when* on the one hand, and discourse adverbials such as *then*, *otherwise*, *nevertheless*, and *instead* on the other hand. It is the latter group, namely discourse adverbials, that, according to Webber et al. (2003), should be considered as anaphors in very much the same way as other anaphoric expressions such as definite descriptions and pronouns.

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.

2. The Data

The British National Corpus (BNC; Burnage and Baguley (1996)) served as the data source for the present investigation. The BNC is a 100 million word collection of samples 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 marked, 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 2693 examples of the phrase in (ADJ) contrast suffices to address the theoretical issues most relevant for this paper.

3. Syntactic Properties

This section provides an overview of the various syntactic environments that the phrase *in contrast* can occur in, all of which are attested in the BNC. *In contrast* can appear either with or without an accompanying prepositional phrase, as shown in (1) and (2), respectively.¹ Among these two options, occurrences without a prepositional phrase are much more frequent in the BNC.

- (1) But both are acceptably direct, although the Corrado's steering has pinpoint accuracy. It's a shame, then, that its gearchange is coarse and sloppy. **In contrast**, the Calibra's is light and quick, although the clutch action could be more progressive. A6W(0763)
- (2) Yet this is the first serious attempt to write about the revolution since the heyday of the early 1970s. In contrast to the books written then, The Road to Jaramillo is full of insight into the interactions, communications, thoughts and impressions that make scientific problem solving enlightening. B72(1514)

While in examples (1) and (2) the phrase *in contrast* appears sentence-initially, it can also appear in non-initial position, as in (3) or even as the last phrase in a sentence, as in (4).

- (3) On the other side of the mill, **in contrast**, was a deep high banked muddy trough meandering the last half mile into the Severn estuary. B3J(1927)
- (4) The vegetation of urban commons varies region by region, and so unwittingly contributes to local character **in contrast with most urban landscapes**. B7L(1619)

Sentence-initial occurrences far outrank non-initial occurrences in the BNC. Among non-initial occurrences, sentence-final placement is more frequent than non-final placement. Thus, the placement of *in contrast* seems to be concentrated at the left or right edge of the clause, with strong preference for the left edge.

Finally, sentence-initially *in contrast* can cooccur with other discourse connectives with related meanings such as *however*, as shown in (5).

(5) Like Mozambique and Nicaragua, it is struggling to survive, with its education system in chaos. However, in contrast, successive Sudanese governments have made little attempt to initiate radical social transformation. B12(0566)

All of the cases considered so far represent cases where *in contrast* functions as a prepositional phrase adjunct of the clause that it appears in. However, *in contrast* can also appear in predicative position with the copula *be* or with light verbs such as *stand* in (6).

(6) South Africa stands **in apparent contrast** to the rest of the states considered here. B12(0020)

(6) also provides an example of an adjectival premodifier that can modify *contrast* and that tends to function as an intensifier. Other such modifiers include *profound*, *sharpest*, *strong*, *utter* and *clear*.²

4. Discourse Anaphora

This section will focus on the discourse function of the adverbial phrase *in contrast*. Following Webber et al. (2003), we will argue that it resembles other discourse adverbials such as *then*, *otherwise*, and *nevertheless* in that it crucially involves the notion of discourse anaphora.

Discourse anaphora involves a relation between an anaphor, such as a pronoun or a temporal adverbial, and an antecedent that is present in the previous discourse or that can be inferred from it. In the case of pronouns, antecedents are typically NPs, while the antecedents of temporal adverbials can be time-denoting expressions, such as dates, events or states of affairs. For pronouns, discourse anaphora can either involve coreference or more indirect referential relations which do not involve identity of reference with a previous discourse entity, but where the anaphor is merely associated with a previously mentioned discourse entity. Such cases of indirect referential relations include cases of bridging, as in (7), where the anaphor, in this case *the receiver* stands in a part-whole-relation to its antecedent – in this case *a phone*.

(7) Myra darted to **a phone** and picked up **the receiver**. (Webber et al. (2003), p. 555)

Other-anaphora (Bierner and Webber (2000) Bierner (2001), Modjeska (2002)), as in (8), provides another instance of such an indirect referential relation.

(8) Sue grabbed one phone, as Tom darted to **the other phone**. (Webber et al. (2003), p. 555)

Here the referent of *the other phone* can be inferred from the antecedent *one phone*. The referential relation between the anaphor and the antecedent is not one of identity of reference. Rather, the referents of the antecedent and the anaphor together constitute the set of phones owned by Sue and Tom.

It is indicative of the anaphoric character of *in contrast* that it licenses *other*-anaphora in the same way, as shown in (9).

(9) He retired to Hampshire and died in 1832 at the age of 76. The MCC continue to care for his grave. In contrast, on the other side of the same cemetery is the grave of the Burgess family, where the ashes of the spy Guy Burgess, who died in Russia, were placed. BM4(0772)

Note that (9) is not an isolated case. (3) shown above provides another example of this kind.

A second piece of evidence in support of the anaphoric properties of *in contrast* concerns ellipsis, as shown in (1), repeated below as (10).

¹The codes below these examples and all other examples taken from the BNC refer to the section and sentence number where the examples can be found in the corpus.

²The following modifying adjectives are attested in the BNC: *abrupt, absolute, apparent, appreciable, bizarre, clear, com-*

plete, delightful, direct, distinct, dramatic, elegant, explicit, extreme, fascinating, frightening, further, great, greater, harmonic, harmonious, high, marked, methodological, moderate, profound, pure, real, sad, sharp, sharpest, significant, sorry, strange, stark, strong, subdued, sympathetic, total, unhappy, utter, and welcome. With the possible exception of subdued, all other adjectives function as intensifiers. Given the lexical meaning of contrast and its discourse function, this should come as no surprise.

(10) It's a shame, then, that its gearchange is coarse and sloppy. In contrast, the Calibra's is light and quick, although the clutch action could be more progressive. A6W(0763)

Here the elliptical *the Calibra's* is missing its nominal head, which is provided by the antecedent *gearchange*. Yet another anaphoric effect licensed by *in contrast* arises with respect to the notion of *domain restriction*, previously studied by, among many others, Lewis (1979), Hinrichs (1988) and Hinrichs (1998), and von Fintel (1994).

(11) Few countries have satisfactory legislation on pesticides or the trained manpower to enforce it. In contrast, extensive use of pesticides in Europe, North America and Japan is backed by government legislation or voluntary schemes that provide farmers with detailed advice. B7G(0726)

Note that the domain of the set of countries in the quantified NP *few countries* is subsequently narrowed so as to not include countries in Europe, North America and Japan. It is precisely the explicitly mentioned contrast that leads to this effect.

Webber et al. (2003) observe that identification of the correct antecedent of a definite description such as *the tower* or *this tower* in (12a) or a discourse adverbial such as *otherwise* in (12b) may require reference to abstract discourse objects such as the result of stacking blocks (to form a tower) or the state of not wanting an apple as the logical antecedent of a definite description or of a discourse adverbial.

- (12) a. Stack five blocks on top of one another. Now close your eyes and try knocking {the tower, this tower} over with your nose. (Webber et al. (2003), p. 552)
 - b. Do you want an apple? **Otherwise** you can have a pear. (Webber et al. (2003), p. 552)

Notice that the same kind of inference is required for *contrast* in example (13), providing further evidence for the anaphoric nature of this discourse connective.

(13) Jack's heart lurched as he saw the ambulances and the busy, functional building and he immediately forgot everything they had been saying. "I'll ask where he is," said Jamie Shepherd as they walked towards the reception desk. **In contrast to the outside**, the area was softly carpeted, softly lit, as if illness and death had to be cushioned away, made to look as if they didn't exist. BPD(0200)

The referent of *outside* in (13) is never explicitly mentioned. Rather, *outside* refers back to the entire scene described before.

Another type of inference that is sometimes necessitated by the *in contrast* connective concerns the operation of complementarity of reference as in (14).³

(14) Other speed-reducing devices may be added, such as regular shifts in the axis of the road, together with changes in the profile in the form of ramps and speed humps (Figure 4.3). Narrowings that allow a cycle to pass but not two cars are frequently added, often reinforced by the placement of trees, planters and street furniture. In contrast to the flowing design of fast roads, design elements are angular and of pedestrianscale, typified by low-level lamp posts which avoid the "sea of light" provided by high poles in traffic streets (Figure 4.4). C8F(0297)

In this text, which is on the topic of child safety, roads are never explicitly mentioned. Rather the concept of slow neighborhood roads can only be inferred from the description. The first explicit mention of the term *road* then refers to the opposite term *fast roads*.

Comparison of *in contrast* with personal pronouns yields yet another similarity with other anaphoric expressions. Like with personal pronouns, the antecedent of *in contrast* can either occur across sentences, as in all of the examples considered so far, or it can occur intrasententially, as in example (15).

(15) **In contrast to his predecessors** who worked at all hours of the day Macmillan tended to keep office hours. B0H(0476)

Another property that distinguishes anaphoric discourse adverbials from structural connectives in the sense of Webber et al. (2003), i.e. coordinating and subordinating conjunctions, concerns the type of dependencies that the arguments of the types of connectives can enter into. While structural connectives only allow non-crossing adjacent material as their arguments, discourse adverbials may involve crossing dependencies among non-adjacent material – just like other anaphoric expressions. e.g. pronouns and definite descriptions. Figure 1 and Figure 2 show this type of crossing dependency for *in contrast* both intrasententially and across sentence boundaries. While Figure 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.

5. Compositional Semantics

This section will try to develop a logical representation for *in contrast* that does justice to its anaphoric properties and to the lexical semantics of the lexeme *contrast*. As will become clear in the course of this discussion, this task is far from trivial. The following remarks should, thus, not be taken as a fully worked-out proposal, but rather as an attempt to point out a set of crucial properties that a fully worked-out account needs to take into consideration.

The discourse properties of *in contrast* are not just of interest from a purely theoretical perspective. Teufel and Moens (2002) and Siddharthan and Teufel (2007) have pointed out that phrases like *in contrast* are typical cases of what they call *argumentative zoning* and are highly relevant for information extraction and text summarization. In particular, for tasks like abstracting scientific papers and detecting novel

³In order to avoid possible confusion: the reference to Figure 4.3 and Figure 4.4 are part of the quoted material from the BNC. They do not refer to actual fi gures contained in the present paper.

[In contrast] [$_{arg1}$ to his predecessors who worked at all hours of the day] [$_{arg2}$ Macmillan tended to keep office hours].

Figure 1: Crossing Intrasentential Dependency

It's a shame, then, that $[arg_1]$ its gearchange is coarse and sloppy]. [In contrast], $[arg_2]$ the Calibra's is light and quick], although the clutch action could be more progressive].

Figure 2: Crossing Intersentential Dependency

claims compared to the previous literature, phrases like *in contrast* can provide important cues.

In the previous section we established at some length that *in contrast* shares with other discourse adverbials its anaphoric behavior. This naturally raises the question whether the semantics that has been proposed for this class of expressions can be naturally generalized to the semantics of *in contrast*.

Following earlier proposals by Hinrichs (1986) and Kamp and Reyle (1993), Webber et al. (2003) assume that the semantics of discourse adverbials such as *then* involves an anaphoric relation between two events. For example, the two clauses in (16) refer to individual events, which are put in the *sequence*-relation by the adverbial *then*.

- (16) a. [$_{e1}$ Remove dish to a wire rack to cool for one hour]. Then [$_{e2}$ refrigerate several hours or overnight].
 - b. sequence(e1,e2)

A straightforward generalization of this approach to the semantics of *in contrast* would involve positing a semantic representation for examples such as (17a) as in (17b).

- (17) a. According to the buyers' guide produced by estate agents Savills, who specialize in such properties, [e1 the price of prime properties in Yorkshire has risen by more than 130 per cent over the last three years.] In contrast, [e2 the price of average properties, those on which the Halifax has lent mortgages, has risen by only about 85 per cent.] A5T(167)
 - b. in-contrast(e1,e2)

There are at least two difficulties associated with modelling the semantics of *in contrast* as a two-place relation between events: The scope of the two arguments of the *contrast* relation often extends beyond descriptions of individual events, as illustrated by examples such as (13), where the contrast involves sets of events and states of affairs. Thus, at the very least, one would have to generalize the semantics of *in contrast* to relations between sets of events and states of affairs, with relations between single events or states of affairs as a special case.⁴ However, it is difficult to see how such a modified representation could be suitably generalized to adequately model examples as in (18).

(18) The Holsteins also tend to have much more white in the coat so that the white areas predominate and they could almost be described as white-and-blacks in contrast to the black-and-white Friesian type. B0K(0438)

(18) explicitly contrasts two sets of individuals, cows in this case, rather than events or states of affairs. One could, of course, argue that the in contrast relation is simply polymorphic, referring either to relations between sets of events or states of affairs or to relations between (sets of) individuals or other entities such as locations, as in (13). Support for such a position could be derived from the fact that there are two syntactic variants of the in contrast connective: one with and one without a postmodifying prepositional phrase. The latter could then be interpreted as involving a relation between sets of events and/or states of affairs, and the former with relations between entities of various sorts, e.g. individuals, locations, times, etc. However, there are two shortcomings of such an account: (i) Everything else being equal, one would prefer a unified account of in contrast and thereby of the two syntactic constructions, and, more importantly, (ii) an account of in contrast without a postmodifying prepositional phrase that merely posits a relation between sets of events and/or states of affairs misses the fact that this construction also focuses on specific participants of the events and/or states of affairs as opposites. (17), for example, focuses on the differences between the price of prime versus average properties. Once one recognizes such contrast pairs for the in contrast construction without a postmodifying prepositional phrase, then a unified analysis is starting to emerge. This analyis can be illustrated by the logical formulas in (19a) and (19b) for examples (17) and (18), respectively.5

⁴Webber et al. (2003) already noted that related discourse connectives like *but* can take multiple sentences as antecedents.

⁵The hyphenated terms in the formulas (19a), (19b), and (20b) are intended to represent translations of the substrings of words in the corresponding sentences into a suitable language of fi rst-order or higher-order logic. Since the focus of the present paper is on the semantics of *in contrast* as such, this simplification seems permissible.

- (19) a. in-contrast((the-price-of-prime-properties-in-Yorkshire, λx [According-to-the-buyers'-guideproduced-by-estate-agents-Savills,-who-specialize-in-such-properties, x has-risen-by-more-than-130-per-cent-over-the-last-three-years]), (the-price-of-average-properties,-those-on-which-the-Halifax-has-lent-mortgages, λx [x has-risen-by-morethan-130-per-cent-over-the-last-three-years])).
 - b. in-contrast((The-Holsteins, λx [x also-tend-tohave-much-more-white-in-the-coat-so-thatthe-white-areas-predominate-and-they-couldalmost-be-described-as-white-and-blacks]), (the-black-and-white-Friesian-type, $\lambda x \neg$ [x alsotend-to-have-much-more-white-in-the-coat-sothat-the-white-areas-predominate-and-they-couldalmost-be-described-as-white-and-blacks])).

In (19), in-contrast is modelled as a two-place relation. Each argument consists of an ordered pair of expressions: the first members of each argument represent the contrasted items, and the second members represent their respective properties which distinguish the items. These properties are represented as lambda-expressions that denote open propositions. Notice also that the open propositions in (19b) represent contradictory properties. The open proposition in the first argument of (19b) is explicitly mentioned in discourse, while its negation, the open proposition in the second argument, is not. Rather, this open proposition can be inferred, given the semantics of contrast. It turns out that the presence of such implicit arguments is what distinguishes the in contrast to construction from the in contrast construction. In the majority of cases of the former construction found in the BNC, the second argument is left implicit and needs to be inferred from the context. In other cases, as in (15) - repeated below as (20a) - both contrasting properties are explicitly mentioned. Hence, the semantics of (20a) can be rendered as in (20b).

- (20) a. **In contrast to his predecessors** who worked at all hours of the day Macmillan tended to keep office hours. B0H(0476)
 - b. in-contrast((Macmillan's-predecessors, λx [x worked-at-all-hours-of-the-day]), (Macmillan, λx [x tended-to-keep-office-hours]))

The account of *in contrast* which has been illustrated by the formulas in (19) and (20b) has two attractive properties: (i) from a theoretical perspective, it provides a unified analysis of the *in contrast* construction with and without a postmodifying prepositional phrase; (ii) by separating out the contrast pairs (as the first members of each argument pair) from their contrasting properties, it provides a transparent representation for applications such as information extraction and text summarization, which require tracking discourse entities and their relevant properties.

Finally, it is worth reviewing the proposed analysis in light of the generalization put forth by Webber (2006) and by Lee et al. (2006), namely that discourse connectives always denote two-place relations. The semantics of *in contrast* proposed in this section is consistent with this hypothesis since it assumes a two-place relation. However, notice that each of the two arguments is further structured into a contrast item and a contrast property. It is this highly structured character of the *in-contrast* relation that distinguishes this discourse connective from the much simpler two-place relations denoted by coordinating and subordinating conjunctions. The latter simply denote relations between events and/or states of affairs, namely those denoted by the two conjunct clauses. The semantics proposed for *in-contrast*, thus, provides further evidence for the distinction between coordinating and subordinating conjunctions and discourse adverbials that has been put forth by Webber et al. (2003).

6. Conclusion and Future Work

This paper has presented a corpus-based study of the discourse connective *in contrast*. The corpus data were drawn from the British National Corpus (BNC) and were analyzed at the levels of syntax, discourse structure, and compositional semantics. Following Webber et al. (2003), the paper argues that *in contrast* crucially involves discourse anaphora and, thus, resembles other discourse adverbials such as *then*, *otherwise*, and *nevertheless*. The compositional semantics proposed for other discourse connectives, however, does not straightforwardly generalize to *in contrast*, for which the notions of contrast pairs and contrast properties are essential.

In future work we plan to consider a wider range of contrast relations in discourse such as *by comparison, contrary to* and *on the other hand* 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 figure prominently in the compositional semantics of *in contrast*, by means of machine learning techniques. Here we expect that elliptical expressions, *other*- anaphora, and syntactic parallelism will provide important cues.

7. Acknowledgements

This research was supported by a grant from the German Research Foundation (DFG) as part of the special research center *Linguistic Data Structures* (SFB 441) at the University of Tübingen. The authors would like to thank Michael White for extensive and extremely helpful comments on an earlier version of this paper, Crystal Nakatsu for fruitful discussions, and Yannick Versley for providing software support for data extraction from the BNC. We are also grateful to two anonymous reviewers of the LREC 2008 program committee for their feedback.

8. References

- Gann Bierner and Bonnie Webber. 2000. Inference through Alternative-Set Semantics. *Journal of Language and Computation*, 1(2):259–274.
- Gann Bierner. 2001. *Alternative Phrases: Theoretical Analysis and Practical Application*. Ph.D. thesis, University of Edinburgh.
- Gavin Burnage and Glynis Baguley. 1996. The British National Corpus. *Library and Information Briefings*, 65.

- Erhard Hinrichs. 1986. Temporal Anaphora in Discourses of English. *Linguistics and Philosophy*, 9(1):63 82.
- Erhard Hinrichs. 1988. Tense, Quantifiers and Contexts. *Computational Linguistics*, 14(2):3–15.
- Erhard Hinrichs. 1998. Interpreting Questions in Discourse. In C. Smith, editor, *Proceedings of the Workshop on the Structure of Non-Narrative Text*, Occasional Papers in Linguistics, Department of Linguistics, Austin, Texas.
- Hans Kamp and Uwe Reyle. 1993. From Discourse to Logic. Kluwer, Dordrecht.
- Alan Lee, Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, and Bonnie Webber. 2006. Complexity of Dependencies in Discourse: Are Dependencies in Discourse More Complex Than in Syntax? In Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories, Prague, Czech Republic.
- David Lewis. 1979. Scorekeeping in a Language Game. Journal of Philosophical Logic, 8:339–359.
- Natalia N. Modjeska. 2002. Lexical and grammatical role constraints in resolving *other*-anaphora. In *Proceedings* of the Discourse Anaphora and Anaphora Resolution Colloquium, pages 129–134, Lisbon, Portugal.
- Advaith Siddharthan and Simone Teufel. 2007. Whose idea was this, and why does it matter? Attributing scientific work to citations. In *Proceedings of NAACL/HLT-07*, Rochester, New York.
- Simone Teufel and Marc Moens. 2002. Summarizing Scientific Articles – Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28(4).
- Kai von Fintel. 1994. *Restrictions on Quantifier Domains*. Ph.D. thesis, University of Massachussetts at Amherst.
- Bonnie Webber, Aravind Joshi, Matthew Stone, and Alistair Knott. 2003. Anaphora and Discourse Structure. *Computational Linguistics*, 29(4):545 – 587.
- Bonnie Webber. 2006. Accounting for Discourse Relations: Constituency and Dependency. In M. Butt, M. Dalrymple, and T. King, editors, *Intelligent Linguistic Architectures*, pages 339 – 360. CSLI Publications, Stanford.