DASHES AS TYPOGRAPHICAL CUES FOR THE INFORMATION STRUCTURE

(Extended Abstract)[*]

Bilge Say and Varol Akman
Department of Computer Engineering and Information Science
Faculty of Engineering, Bilkent University
Bilkent, Ankara 06533, Turkey
Phone: [90] (312) 266–4133 (secretary)
Fax: [90] (312) 266–4126
{say,akman}@cs.bilkent.edu.tr

Conference Topic: Information-based approaches to syntax, semantics, and pragmatics of natural language

[*] Accepted for presentation at ITALLC 98. Please send all correspondence to the second author.

We take em-dash as our sample punctuation mark and examine its usage from a discourse perspective, using sentences from well-known corpora. We particularly comment on how dashes can give hints on information structure, focus, and anaphora. Throughout the paper Discourse Representation Theory is used as a framework.

Keywords: Punctuation, Discourse, Discourse Representation Theory, Information Structure

1 Introduction

To the initial onlooker, punctuation marks and topic/focus structure of an orthographic sentence seem to be unrelated. However, recent linguistic and computational research in punctuation [3, 13, 21] have suggested that there may be interesting consequences of punctuation for the semantics of discourse. Nunberg [17] has suggested that the contributions of punctuation can be studied by means of a linguistic characterization of underlying categories and attached constraints. Dale [8] has observed that many uses of certain marks (e.g., comma, colon, semicolon, dash(es), parentheses) act as signals of discourse structure. In this paper, we concentrate on em-dash as our sample punctuation mark and investigate its effects on information structure, focus, and anaphora.

We take Discourse Representation Theory (DRT) [15], an influential theory dealing with various discourse related phenomena, as our framework. In particular, Asher's extension [2] of DRT with Segmented Discourse Representation Structures (SDRSs) proves valuable for us. SDRSs provide various mechanisms to represent discourse structure and constraints on those representations for the resolution of abstract entity anaphora.[1] SDRT has been used in a way similar to ours for a different problem: translation of a syntactically-complex, informationally dense, hypotactical language (German) into a less complex, paratactical language (Norwegian) [10]. We see SDRT as a valuable framework for exploring problems of punctuation in which informational characteristics of written text are central [20, 19].

In Section 2, we examine how dashes act as typographical cues for the information structure in written English. We then examine how anaphora resolution is

[1] Abstract entity anaphors are those anaphors that refer back to propositions, facts, or eventualities that are not precisely definable with causal, temporal, or spatial properties [2, p. 2].

affected between dash interpolated text adjuncts and other text. In Section 3, we give the necessary extensions to SDRT to be able to model the observed data.

2 Constraints on Discourse Related Phenomena

The following observations are due to examining around 500 randomly selected sentences with dash interpolations. The sentences come from the *Wall Street Journal* [1], the British National Corpus (BNC) [5], and SUSANNE [18]. Depending on their origin, they are marked **W**, **B**, and **S**, respectively, in the following.

2.1 Constraints on Information Structure and Focus

One observation is that dash interpolations do cue a subset of discourse relations quite reliably. By discourse relations, we mean those bindings that relate two units of text by means of coherence and rhetorical effects [16]. In Table 1, the discourse relations that are paired with the dash interpolations are shown. Note that although there is a distribution pattern, these relations are due to the semantics and context of the sentences; thus their distribution could be slightly different for another interpreter who tries to classify them into the same set of relations. We tried to stick to the relations used by Asher [2], as we used mostly the same kind of texts (i.e., "news articles").

The row denoting other (nonrelational) usages in Table 1 consists of corpus-specific reference mechanisms, title introduction, list introduction using dashes, or one-off usages such as introducing quoted sentences. As can generally be seen, the distribution of discourse relations in dash interpolated sentences is not completely ad hoc and is worthy of special consideration. (Examples for each relation can be found in the Appendix A.) Indeed, 56% of the relations are in the categories of Elaboration, Commentary, and Apposition.[2]

The second observation that relates to an interesting use of dash interpolation is when it is used to denote *focus*—in a combination of the informational sense

[2] The reason we treat apposition as a kind of discourse relation when used in conjunction with dashes is that it is usually a special case of apposition with emphasis. This effect is more clearly seen if we think of the possibility of substituting another mark (or marks) in place of the dash(es). In the case of apposition, dashes can mostly be replaced by commas but with some loss of emphasis in most of the occurrences.

Discourse Relations	**No. of Sentences**	
Elaboration	26	20.8%
Commentary	20	16%
Apposition	24	19.2%
Explanation	12	9.6%
Contrast	6	4.8%
Parallel	3	2.4%
Result	2	1.6%
Instance	3	2.4%
Continuation	2	1.6%
Cause	2	1.6%
Informational Focus only	2	1.6%
Background	1	0.8%
Other (Nonrelational)	22	17.6%
Total	125	100%

Table 1: Distribution of Discourse Relations for Dashes

(used, for example, within the studies of information packaging[3] by Vallduví and Engdahl [23]) and the intonational sense. Informational focus of a sentence is the informative (new) part of a sentence that makes a contribution to a hearer's mental store. Intonational focus, on the other hand, indicates intonational prominence denoted by any constituent that bears a pitch accent. In English, a subset of the informational focus is realized in situ by intonational prominence. Not all accented constituents are parts of the informational focus though; they may be a part of topic/link [12].

Some dashes do not disrupt the syntactic flow of the sentence. In other words, they solely add an element of emphasis. This could indicate an extra level of emphasis on informational prominence, where an intonational focus would already be expected in spoken text (see example (1)), or distinguish what would have been an intonational focus on a lexical word or phrase in speech (see example (2)).

(1) **(W)** Already, the consequences are being felt by other players in the financial markets—even governments.

(2) **(W)** Knowing a tasty—and free—meal when they eat one, the executives gave the chefs a standing ovation.

If the dash interpolation comes at the end of the sentence it is usually more prominent informationwise than its mid-sentence counterparts. This might be due to the fact that it is cognitively more plausible for the human mind to consume the information acquired most recently [22]. Compare (3) with (4) as examples of changing prominence.

(3) **(W)** In addition, the Cray-3 will contain 16 processors—twice as many as the largest current supercomputer.

(4) **(W)** Some of the biggest service-industry exporters—American financial-service companies, for example—have yet to be fully included in our export statistics.

[3] "Information packaging is a structuring of sentences by syntactic, prosodic, or morphological means that arises from the need to meet the communicative demands of a particular context or discourse" [23, p. 460]. We hypothesize that orthographic means might in some ways be helpful for information packaging.

Some styles of writing (such as a brochure by a health organization, as found in the BNC [5]) make repeated use of the above effect and employ dash interpolations for intonational focus to keep a vivid and striking pace throughout the document.

An end-of-sentence dash interpolation might convey key information in the form of a text-phrase that gives out some information otherwise not mentioned overtly in the sentence. In such cases, intonational focus and part of informational focus fall on the dash interpolation (see example (5)).

(5) **(W)** As a result, marketers of faux gems steadily lost space in department stores to more fashionable rivals—cosmetics makers.

Even when they are part of informational focus, mid-sentence dash interpolations can be parenthetical and less prominent than other parts of the sentence. They can give background or extra information and comments that are not necessarily crucial to the understanding of the text-sentence (see examples (6) and (7)).

(6) **(W)** The department said orders for nondurable goods—those intended to last fewer than three years—fell 0.3% in September to $109.73 billion after climbing 0.9% the month before.

(7) **(W)** Still, the restaurant's ever-changing menu of five-course dinners—it supposedly hasn't repeated a meal since opening in 1971—requires constant improvisation.

On the other hand, dash interpolations can also change the perspective of the reader by offering an alternative wording, e.g., (8). Within the dash interpolation, the reader is directed to a different encyclopedic entry in a relevance-theoretic way in that the writer uses the dash interpolation as a means to establish the maximum contextual effect with minimal processing effort for the reader by overriding or strengthening the meaning of the lexical entry it is adjoined to [4, 22].

(8) **(W)** They showed up, but didn't—or couldn't—challenge.

2.2 Constraints on Anaphora

The next question to consider is whether these observations at the discourse level have implications for anaphora resolution. The basic observation is that antecedents within a dash interpolation are less felicitous if the dash interpolation has an adjoining (parenthetical) status and is mid-sentence (except for the case that the antecedents introduced within the dashes form an apposition to the noun phrase before the dash that they are adjoined to). This is not so with conjoining status dash interpolations where other factors (grammatical function, lexical iteration, etc.) function as normal.[4]

See example (9) where the principle seems to have been violated. Native speakers found *these countries* to be ambiguous as to which countries it included and which it did not.

(9) **(W)** On Asia-Pacific prosperity: "If America can keep up the present situation—her markets open for another 15 years, with adjustments, and Japan can grow and not cut back, and so too, Korea, Taiwan, Hong Kong, Singapore, ASEAN, Australia and New Zealand—then in 15 years, the economies of these countries would be totally restructured to be able to almost sustain growth by themselves." In such an arrangement, "all benefit," he said. "And if the Europeans come in, they benefit too. It's not a zero-sum game."

On the other hand, in (10), “their parents” do not stand as a felicitous canditate for further anaphoric reference, though it stands in the subject position (a strong position to be an anaphoric canditate, though it is already anaphoric in form) from within the dash interpolation.

(10) **(W)** The issue is further complicated because although the organizations represent Korean residents, those residents were largely born and raised in Japan and many speak only Japanese. That they retain Korean citizenship and ties is a reflection of history—their parents were shipped in as laborers during the decades when Japan occupied Korea before World War II—and the discrimination that still faces Koreans in Japanese society.

A complementary hypothesis worth looking at is thus as follows: when a dash interpolation is the rightmost constituent and falls on a lexical text-phrase, any discourse referent introduced in that phrase (see example (5) where the second sentence is made-up by us) is a more salient choice than it would otherwise be, for

[4] In the former case, where there is a parenthetical dash interpolation, other factors of anaphoric reference as depicted by theories such as the centering framework [11] still continue to function. The existence of the parenthetical may serve as a preference or overriding factor.

serving as an antecedent in the next sentence. The resolution of *they* to cosmetic makers seems to be a more felicitous choice unless the context dictates otherwise.

(11) (=5) **(W)** As a result, marketers of faux gems steadily lost space in department stores to more fashionable rivals—cosmetics makers. They are really aggressive.

To be able to exactly determine the saliency changes, a more elaborate and detailed study is needed as dashes are not used in such a widespread way as commas (viz. dashes constitute normally 2% to 5% of all the punctuation marks in a corpus, according to [14]).

3 Extensions to SDRT

In this section, we give three additions to SDRT based on our findings. Note that there have been suggestions on integrating focus and information structure into DRT [7, 9]. Here, we try to concentrate on the implications of dashed sentences. We adopt the basic definitions of DRT and SDRT as they are given in Asher [2]. In DRT, a semantic representation structure called Discourse Representation Structure (DRS) is assigned to a discourse segment by a syntax-driven construction algorithm that proceeds sentence by sentence relative to the context created by the DRS built so far. A DRS contains a set of discourse referents which corresponds to the entities in the discourse, and a set of conditions (quantificational, temporal, etc.) on those referents. The later step of interpretation associates the discourse referents with entities and verifies conditions. In SDRT, there is an additional level of discourse interpretation in the form of Segmented DRSs (SDRSs) which are constructed by default for each orthographic sentence and enter into discourse relations with each other forming a discourse structure.

From now on, we will take an SDRS as corresponding to a text-sentence (an orthographic sentence), a text-clause (clauses or lexical sentences separated by semicolons conjoined at the same level) or a text-phrase (lexical sentences, clauses or phrases separated or delimited by colons, parentheticals or dashes) as justified by the dashes. The definitions of text-sentence, text-clause, and text-phrase are borrowed from Nunberg [17]. To the definition of the construction algorithm [2, pp. 302–304], we add the following: If a, b, c are text-sentences, text-clauses, or text-phrases and α, β, γ are the corresponding DRSs or SDRSs, then a—b—c

(that is, a followed by an em-dash followed by b followed by an em-dash followed by c) will lead to the following conditions:

- If in a—b—c, c is not empty and b is an S or a VP in the lexical grammar, then any discourse referent introduced in β is defeasibly inaccessible to discourse referents introduced in γ and the following text-sentences of the text-segment (a.k.a. Double-Box Constraint).
- If in a—b—c, c is empty and b is an NP, mark the discourse referent (e.g., by underlining) as a salient entry for further antecedent choice.
- If in a—b—c, a and b belong to the same verbal category (V, AUX), show the overriding affect in the conditions by the sign $\triangleright$.
- If the pattern is a—b, that is, c is empty as the dash used is a conjoining one and a and b are text-sentences, then construct SDRSs in the usual way as depicted in Asher [2]. Do not put the Double-Box Constraint in use.

The first modification in the list above is to encode a way to denote that certain discourse referents are not preferable for selection (though they are available). Both in DRT and SDRT, whether a discourse referent is available as an antecedent is strictly defined with openness and availability constraints. However, in sentences such as (12), there should be a way to denote that the discourse referents introduced in the dashed sentence are parenthetical and are not preferred for further selection. In example (12), *he* is resolved to be John, rather than his brother. (Caveat: Some native speakers find this example somewhat forced.)

(12) John—his brother is also an athlete—won the university medal for 3000m easily. He is an ambitious guy.

We choose to use a double-framed box for this purpose (see Figure 1).

The second modification is to make discourse referents such as those in example (13) more prominent than others as they denote special intonational focus. This we denote with an underline (see Figure 2).

(13) **(S)** [simplified] This is the underlying concern along with the lack of time—the shortage of cash. It is an acute problem.

The third modification takes into account the alternative wording effect of dash interpolations such as example (14) (from [4, p. 116]).

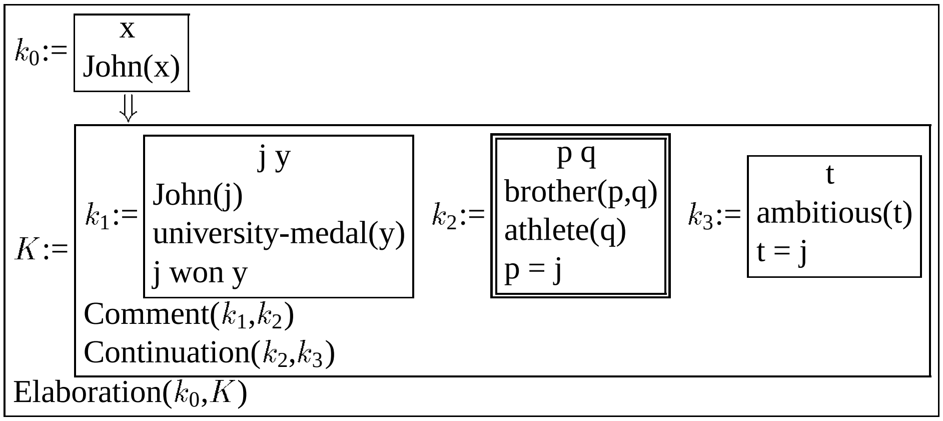

Figure 1: SDRS for example (12)

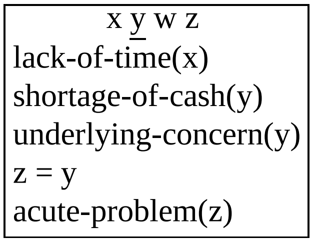

Figure 2: DRS for example (13)

$$\begin{array}{|l|}\hline \qquad\quad \text{u v n e t} \\ \text{u} = ? \\ \text{e} \subseteq \text{t} \\ \text{t} < \text{n} \\ \text{e:run(u)} \triangleright \text{e:sprint(u)} \\ \text{the hill(v)} \\ \text{up(u,v)} \\ \hline \end{array}$$

Figure 3: DRS for example (14)

(14) They ran—sprinted—up the hill.

The ▹ sign in the conditions of Figure 3 shows the overriding affect.

4 Conclusion

A more detailed study is being performed as to whether dashes and other semantically significant punctuation marks can provide defeasible cues (such as the use of discourse particles *but* and *too* in denoting Contrast and Parallel [2, p. 286]) for the discourse interpretation of punctuated orthographic sentences. Such a characterization may also lead to their inclusion in the relevant computational modules such as the anaphora resolution component of an NLP system already accounting for punctuation at a syntactic level [6]. Extending SDRT (and DRT) with suitable triggering conditions is our main project. We see our attempt as beneficial and contributing to the efforts for modifying DRT to cover sentences from corpora.

A Examples of Discourse Relations

Precise definitions of the relations occurring in the following sentences can be found in [2, pp. 299–309]. Below examples are given for each relation that is used in Table 1.

1. *Apposition*

   (S) Liberals and conservatives in both parties—democratic and republican—shall divorce themselves and form two independent parties, George H.

Reama, nationally known labour management expert, said here yesterday.

(B) Of five such founding fathers—Marx, Comte, Spencer, Durkheim and Weber—Marx (1818–83) and Weber (1864–1920) alone held what could be described as “emancipated” views about women.

2. *Background*

   (W) Mr. Quinlan, 30 years old, knew he carried a damaged gene, having lost an eye to the rare tumor when he was only two months old—after his mother had suffered the same fate when she was a baby.

3. *Contrast - Parallel*

   (W) Learning skills, producing something cooperatively, feeling useful, they are no longer dependent—others now depend on them.

4. *Continuation*

   (B) Wallace Arnold (0532–311055) is the accredited coach-tour operator from the UK—a three-day stay at the Hotel Cheyenne for two adults sharing a room ranges from around *$*130–*$*150 per person (additional child *$*65–*$*81).

5. *Commentary*

   (W) But as they hurl fireballs that smolder rather than burn, and relive old duels in the sun, it’s clear that most are there to make their fans cheer again or recapture the camaraderie of seasons past or prove to themselves and their colleagues that they still have it—or something close to it.

6. *Elaboration*

   (W) In late trading, the shares were up a whopping 122 pence ($ 1.93)—a 16.3% gain—to a record 869 pence on very heavy volume of 9.7 million shares.

   (S) The social security pay-roll tax is now 6 per cent—3 per cent on each worker and employer—on the first $ 4,800 of pay per year.

7. *Explanation*

   (B) Gary Cattermole remained unbeaten in the latter match, although it was close—defeating Jim Laxton 21–16 in the third, Ron Covall 21–17 in the third and Joe Murray 21–19 in the third.

8. *Instance*

   (W) In this connection, it is important to note that several members of New York's sitting City Council represent heterogeneous districts that bring together sizable black, Hispanic, and non-Hispanic white populations—Carolyn Maloney's 8th district in northern Manhattan and the south Bronx and Susan Alter's 25th district in Brooklyn, for example.

9. *Result*

   (W) Mr. Steinhardt, who runs about $ 1.7 billion for Steinhardt Partners, made his name as a gunslinging trader, moving in and out of stocks with agility—enriching himself and his investment clients.

   (B) Yesterday, however, American announced that the Stansted–Chicago service will end with the last flight on May 31—putting the jobs of 50 ground staff at risk.

References

[1] ACL/DCI. Association for Computational Linguistics Data Collection Initiative, CD-ROM 1, 1991. Information available from Linguistic Data Consortium on the WWW: http://www.ldc.upenn.edu.

[2] Nicholas Asher. *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht, Netherlands, 1993.

[3] Murat Bayraktar, Bilge Say, and Varol Akman. An Analysis of English Punctuation: The Special Case of Comma. *International Journal of Corpus Linguistics* (in press).

[4] Diane Blakemore. Are Apposition Markers Discourse Markers? *Journal of Linguistics*, 32:325–347, 1996.

[5] BNC. British National Corpus. Information available on the WWW: http://info.ox.ac.uk/bnc/.

[6] Ted Briscoe. The Syntax and Semantics of Punctuation and Its Use in Interpretation. In [13], pages 1–8.

[7] Sophie Cormack. *Focus and Discourse Representation Theory*. PhD thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh, UK, 1992.

[8] Robert Dale. Exploring the Role of Punctuation in the Signalling of Discourse Structure. In *Proceedings of a Workshop on Text Representation and Domain Modelling: Ideas from Linguistics and AI*, pages 110–120. Technical University of Berlin, Berlin, Germany, 1991.

[9] Kurt Eberle. Disambiguation by Information Structure in DRT. In *Proceedings of 16th International Conference on Computational Linguistics (COLING '96)*, pages 334–339, Copenhagen, Denmark, 1996.

[10] Cathrine Fabricius-Hansen. Informational Density: A Problem for Translation and Translation Theory. *Linguistics*, 34:521–565, 1996.

[11] Barbara Grosz, Aravind Joshi and Scott Weinstein. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–225, 1995.

[12] Herman Hendriks. Information Packaging: From Cards to Boxes. In T. Galloway and J. Spence, editors, *Proceedings of Semantics and Linguistics Theory (SALT) VI*, Cornell University, Ithaca, NY, 1996.

[13] Bernard Jones, editor. *Punctuation in Computational Linguistics*, UCSC, Santa Cruz, CA, 1996. SIGPARSE 1996 (Post Conference Workshop of ACL96). Available from Human Communication Research Centre, University of Edinburgh: http://www.cogsci.ed.ac.uk/hcrc/publications/wp-2.html.

[14] Bernard Jones. *What's the Point? A (Computational) Theory of Punctuation*. PhD thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh, UK, 1997.

[15] Hans Kamp and Uwe Reyle. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht, Netherlands, 1993.

[16] William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: A Theory of Text Organization. Technical Report RS-87-190, USC Information

Sciences Institute, University of Southern California, Marina Del Rey, CA, 1987.

[17] Geoffrey Nunberg. *The Linguistics of Punctuation*. Number 18 in CSLI Lecture Notes. CSLI Publications, Stanford, CA, 1990.

[18] Geoffrey Sampson. *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Oxford University Press, Oxford, UK, 1995.

[19] Bilge Say. An Information-Based Approach to Punctuation. In *Proceedings of Fourteenth National Conference on Artificial Intelligence (AAAI '97)*, page 818, Providence, Rhode Island, 1997.

[20] Bilge Say and Varol Akman. An Information-Based Treatment of Punctuation in Discourse Representation Theory. In *Proceedings of Second International Conference on Mathematical Linguistics* (ICML), pages 93–95, Tarragona, Spain, 1996.

[21] Bilge Say and Varol Akman. Current Approaches to Punctuation in Computational Linguistics. *Computers and the Humanities*, 30(6): 457–469, 1997.

[22] Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition.* Blackwell, Oxford, UK, 1986.

[23] Enric Vallduví and Elisabet Engdahl. The Linguistic Realization of Information Packaging. *Linguistics*, 34: 459–519, 1996.