An Information-Based Approach to Punctuation

Bilge Say
Dept. of Computer Engineering and Information Science
Bilkent University
Bilkent, Ankara 06533, Turkey
Email: say@cs.bilkent.edu.tr

Punctuation marks have a special importance in bringing out the meaning of a text. There has been recent computational work concentrating on punctuation marks in Natural Language Processing (NLP) mostly following Nunberg's pioneering work (Nunberg 1990), in which he bridged the gap between descriptive linguistic treatments of actual usage of punctuation and prescriptive accounts, by putting down the features of a "text grammar" for the orthographic sentence. Several grammars for syntactic parsing incorporating punctuation were then shown by NLP researchers to reduce parse failures and ambiguities in parsing (Briscoe 1996). Nunberg's approach to presenting punctuation (and other formatting devices) was partially incorporated into Natural Language Generation systems. However, little has been done on how punctuation marks bring semantic and discourse-based cues to the text and whether those cues can be exploited computationally. The aim of this thesis is to analyze, in an information-based framework, the semantic and discourse aspects of punctuation, drawing computational implications for NLP systems. This will not only enable NLP software writers to make use of the punctuation marks effectively but also may reveal interesting linguistic phenomena in conjunction with punctuation marks.

Discourse Representation Theory (DRT) (Kamp and Reyle 1993) is taken as the theoretical framework of the thesis because DRT is a dynamic, information-based theory dealing with various semantic and discourse related phenomena. In particular, Asher's (Asher 1993) extension to DRT, viz. Segmented Discourse Representation Theory (SDRT) with constituents called Segmented Discourse Representation Structures (SDRSs), proves valuable as SDRSs provide various devices to represent discourse structure and constraints on those representations for the resolution of abstract anaphora. Included within the definition of SDRSs are precise definitions (in terms of logic) of discourse relations and a defeasible logic for inferencing.

So far, a preliminary study has been done to show how pieces of discourse containing sentences with punctuation can affect discourse phenomena such as anaphora resolution or behave contrary to the expectations of DRT or SDRT (Say and Akman 1996). A more detailed study based on observations from several computerized English corpora is being conducted on the usage of dashes (Say and Akman 1997). Sentences with dashes tend the favor certain discourse relations more than the others in specific ways such as their parenthetical usage. Moreover, dashed sentences have characteristic features in terms of anaphora resolution and determination of focus and information structure.

Future work will involve similar corpus-based studies of several punctuation marks (semicolon, colon and parentheses) and incorporating the findings into a model for semantic and discourse-wise implications of punctuation.

Acknowledgments

This work is being carried out under the supervision of Varol Akman (Bilkent University). Many thanks to Akman, Ted Briscoe (Cambridge University), AAAI, and the Scientific and Technical Research Council of Turkey (TÜBİTAK) for support.

References

Asher, N. 1993. *Reference to Abstract Objects in Discourse*. Dordrecht, Netherlands: Kluwer.

Briscoe, T. 1996. The Syntax and Semantics of Punctuation and Its Use in Interpretation. In *Punctuation in Computational Linguistics*, 1–8. UCSC, Santa Cruz, CA: SIGPARSE 1996 (Post Conference Workshop of ACL96). http://www.cogsci.ed.ac.uk/hcrc/publications/wp-2.html.

Kamp, H., and Reyle, U. 1993. *From Discourse to Logic*. Dordrecht, Netherlands: Kluwer.

Nunberg, G. 1990. *The Linguistics of Punctuation*. Number 18 in CSLI Lecture Notes. Stanford, CA: Stanford University Press.

Say, B., and Akman, V. 1996. Information-Based Aspects of Punctuation. In *Punctuation in Computational Linguistics*, 49–56. UCSC, Santa Cruz, CA: SIGPARSE 1996 (Post Conference Workshop of ACL96). http://www.cogsci.ed.ac.uk/hcrc/publications/wp-2.html.

Say, B., and Akman, V. 1997. A Case for Punctuation within Discourse Representation. Manuscript.