REFERENTIAL CHOICE AS A PROBABILISTIC MULTI-FACTORIAL PROCESS

Andrej A. Kibrik[1], Grigoriy B. Dobrov[2], Natalia V. Loukachevitch[2], Dmitrij A. Zalmanov
aakibrik@gmail.com, wslc@rambler.ru, louk@mail.cir.ru, dm.zalmanov@gmail.com

[1]Institute of Linguistics, Russian Academy of Sciences (Russia)
[2]Moscow State University (Russia)

1. Referential choice. As people talk or write, they recurrently mention the same entities (usually called *referents* in linguistics) over and over in discourse. Repeated reference constitutes one of the founding properties of natural discourse. As one mentions a referent, one faces a choice: whether to use a full lexical account of the referent (full noun phrase), or use a semantically reduced referential device, such as a third person pronoun. Consider a natural example from written discourse:

Tandy said consumer electronics sales at its Radio Shack stores have been slow, partly because a lack of hot, new products. Radio Shack continues to be lackluster, said Dennis Telzrow, analyst with Eppler, Guerin Turner in Dallas. He said Tandy has done a decent job increasing sales by manufacturing computers for others and expanding sales of its Grid Systems Corp. subsidiary, which sells computers to bigger businesses, but it 's not enough to offset the problems at Radio Shack. Sales at Radio Shack stores open more than a year grew only 2 % in the quarter from a year earlier, he said. As a result, Mr. Telzrow said he cut his fiscal 1990 per - share earnings estimate for Tandy to $ 4. 05 from $ 4. 2.

In this excerpt from Wall Street Journal three referents are recurrently mentioned: the company Tandy Corp. (five times), Radio Shack stores (four times), and the individual Dennis Telzrow (six times). Out of five mentions of Tandy, three are by means of a full noun phrase (specifically, proper name), and two by the third person pronoun *it*. Dennis Telzrow is mentioned by a full NP twice, and four times by the third person pronoun *he*. Every time a speaker/writer needs to mention a referent, s/he makes a *choice* between a full and a reduced referential device. What is this choice conditioned by? As is proposed in Kibrik 1999 and a number of other publications, referential choice (**RC**) immediately depends on the level of referent activation in speaker's working memory, called *activation score*. Referent's current activation score, in turn, depends on a number of *activation factors*, that is, properties of the referent and of the discourse context (see below). In the study Grüning and Kibrik 2005 referential choice was modeled with neural networks, a machine learning algorithm, that predicted the actual choice in discourse with a fair level of accuracy.

2. The data. In contrast to Kibrik 1999 and Grüning and Kibrik 2005, the present study is based on a relatively large data set. This is a corpus of texts from Wall Street Journal, such as the one cited above. The corpus contains about 3500 referential devices that are subject to RC. Personal pronouns constitute about 28% of all referential devices, while full NPs account for 72% (including 44% proper names and 28% definite descriptions). Our corpus is based on a prior product, so-called RST Discourse Treebank (see Carlson et al. 2003). That treebank contains annotation of discourse organization in terms of Rhetorical Structure Theory (see Mann and Thompson 1988). The same set of texts has been annotated for a number of candidate activation factors, in accordance with the scheme developed by Krasavina and Chiarcos (2007). Annotation was performed with the program MMAX-2 (http://mmax2.sourceforge.net/), designed specifically for referential corpora.

3. The approach. We attempt to discover dependencies between the set of activation factors and the resultant RC. The following factors have been used in the study:

Properties of the referent:

- referentiality: new in discourse vs. previously mentioned
- animacy: human vs. inanimate
- protagonisthood

Properties of the projected referential device

- plane of discourse: main line vs. quoted speech
- phrase type: noun phrase vs. prepositional phrase
- grammatical role: subject vs. direct object vs. indirect object

Properties of the antecedent (=prior referent mention):

- plane of discourse: main line vs. quoted speech

- phrase type: noun phrase vs. prepositional phrase
- grammatical role: subject vs. direct object vs. indirect object
- referential form: personal pronoun vs. another type of pronouns (reflexive, relative...) vs. proper name vs. *the*-NP vs. possessive NP vs. demonstrative NP

Distance from the projected referential device to the antecedent:

- distance in the number of words
- distance in the number of markables
- distance in clauses along the linear discourse structure
- distance in elementary discourse units along the hierarchical discourse structure.

Statistical modeling of RC has been performed, with the use of the Weka system – a collection of machine learning algorithms for data mining tasks (see Hall et al. 2009). Those machine learning algorithms have been chosen that provide more easily interpretable results, including two logical algorithms: the decision tree algorithm C4.5 and the decision rule algorithm JRip. Also, the logistic regression algorithm has been used, as it is more reliable than the logical algorithms and allows one to obtain probabilistic estimates for the referential options (see below).

4. Results. When all factors are taken into account, the accuracy of prediction of the basic RC (between full and reduced devices), provided by the three algorithms, falls in the range between 85% and 87%. Still being at an early stage in this project, we evaluate this result as satisfactory. When an attempt is made to predict the three-fold choice between pronouns, proper names and definite descriptions, the accuracy of prediction drops to about 75%. This is no surprise, as the set of factors are specifically designed to target the basic RC, and the further choice between proper names and definite NPs does not have to necessarily be accounted for by the same set of factors.

5. Discussion. RC is among the most fundamental phenomena involved in natural language use. In discourse, about every third word is in this or that way related to RC, so understanding RC accounts for an important share of our understanding of language. RC belongs to the class of probabilistic choices: there are instances in which the use of a full vs. reduced referential device is determined, but there are also intermediate instances in which both options are appropriate (Kibrik 1999). So RC is not always categorical; sometimes the outcome of a realistic model of RC should predict a free alternation of a full and a reduced device, or a soft preference of one device over the other.

Provided the above-mentioned probabilistic character of RC, it would be highly unnatural to expect the 100% accuracy of this model. In fact, our model embraces the fluid character of natural RC. Under certain combinations of the values of activation factors, both a full and a reduced referential device must be allowable. Such architecture of the statistical model approaches the cognitive nature of RC, namely the activation score component that serves as a resultant force of all activation factors, involved in every particular instance. Particularly useful in this sense is the logistic regression algorithm: it provides a numerical weight in each case that can be reanalyzed as the probability of using a personal pronoun at the given point in discourse.

The multi-factorial probabilistic methodology developed in this study for modeling RC can possibly be extended to a wide range of other linguistic processes, as well as to non-linguistic types of human purposeful behavior.

This study was supported by grant #09-06-00390 from the Russian Foundation for Basic Research.

Carlson, Lynn D., Daniel Marcu and Mary Ellen Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. Current directions in discourse and dialogue, Kluwer Academic Publishers, 85-112.

Grüning, André, and Andrej A. Kibrik. 2005. Modeling referential choice in discourse: A cognitive calculative approach and a neural networks approach. In: António Branco, Tony McEnery and Ruslan Mitkov (eds.). "Anaphora Processing: Linguistic, Cognitive and Computational Modelling". Amsterdam: Benjamins, 163-198.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. 2009. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1 (http://www.cs.waikato.ac.nz/ml/weka/).

Kibrik, Andrej A. 1999. Reference and working memory: Cognitive inferences from discourse observation. In: Discourse Studies in Cognitive Linguistics. Ed. by K. van Hoek, A. A. Kibrik and A. Noordman. Amsterdam and Philadelphia, John Benjamins, 29-52.

Krasavina, Olga, and Christian Chiarcos, Ch. 2007. PoCoS - Potsdam Coreference Scheme. In: Proceedings of the Linguistic Annotation Workshop (LAW). June 2007, Prague, Czech Republic. Association for Computational Linguistics. p. 156–163.

Mann, William C. and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organisation. In Text 8(3): 243–281.