Referential choice: A cognitively based modeling study

Andrej A. Kibrik (aakibrik@gmail.com)
Institute of Linguistics, Russian Academy of Sciences, and Lomonosov Moscow State University
B. Kislovsky per., dom 1, 125009 Moscow, Russia

Mariya V. Khudyakova (mariya.kh@gmail.com)
National Research University Higher School of Economics,
20 Myasnitskaya Ulitsa, Moscow 101000, Russia

Grigory B. Dobrov (wslcdg@gmail.com)
Trafika
Ozerkovskaja nab. 50-1, 340, Moscow, Russia

Anastasia S. Linnik (anastasia.linnik@gmail.com)
University of Potsdam, Karl-Liebknechtstrasse 24-25, Potsdam, Germany

Abstract

Referential choice depends on referent activation in working memory, and the latter is determined by multiple activation factors. We explore a corpus of texts in which potential activation factors were annotated. Methods of machine learning have made it possible to predict referential choices in the corpus with high accuracy, close to 90% in the case of the basic referential choice between full and reduced referential devices. We propose that many of the instances of divergence between the human and algorithmic production of referential devices result from intermediate referent activation. In an experimental study, we found that human participants usually accept the referential option predicted by the algorithm.

Keywords: referential choice; reference and cognition; multi-factorial analysis; machine learning; corpus annotation; non-categorical referential choice.

1. Introduction: Referential choice

What guides a speaker/writer in deciding which linguistic expression to use when s/he wishes to mention a certain referent? In addressing this question, we rely on a cognitive theory of reference developed in Kibrik (2011) and based on a number of earlier linguistic studies, such as Chafe (1994), Fox (1987), and Givón (1983), as well as on relevant knowledge from cognitive psychology and neuroscience. The process of mentioning referents in discourse, known as reference, is the linguistic manifestation of the general process of attention to entities. When a speaker's decision to attend (mention) a referent is in place, a related phenomenon of referential choice comes into play, guided by the cognitive component of working memory. In the cognitive system, attention controls working memory (cf. Awh et al. 2006): what is attended at moment n is activated in working memory at moment n+1. In the discourse structure, this general relationship between attention and working memory is manifested by the relationship between antecedents and referential expressions. Discourse time can be measured in elementary discourse units (EDUs), roughly equaling clauses: if a referent is attended (mentioned) in EDU n, its maximal activation should be expected in EDU n+1.

Referent's degree of activation in the working memory is immediately responsible for the most basic referential choice. If a referent's activation is high, a lexically reduced referential device may be used, such as a pronoun. If a referent's activation is low, a lexically full referential device is used. Lexically full referential devices are extremely diverse. Two major types must be distinguished: proper names and descriptions, that is, common nouns with or without modifiers. There are many further levels of granularity in this taxonomy. For example, a description may contain a demonstrative pronoun, a possessive pronoun, an adjectival attribute, etc. In our study we are primarily interested in the two upper levels of the taxonomy, that is, the cognitively motivated choice between pronouns and full NPs and, within the latter, between proper names and descriptions.

2. Multi-factorial approach

Referent's activation, responsible for the speaker's choice between reduced and full referential devices, depends on a variety of activation factors. As is shown in Figure 2, activation factors fall into two major groups, including those related to the referent's internal properties, such as animacy, and those related to discourse context. Context factors may be further divided into those associated with the properties of the projected referential expression, with the properties of the antecedent, and with the discourse distance between the projected expression and the antecedent. Referential choice is an inherently multi-factorial process, and it is impractical to try to reduce it to one or few factors.

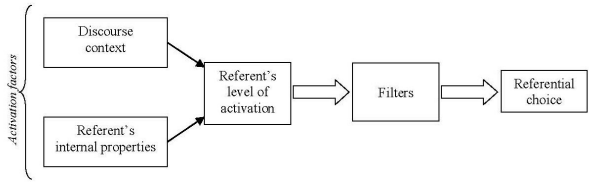

Figure 1: The cognitive multi-factorial model of basic referential choice

As was argued in the line of studies begun by Fox (1987) and continued by Kibrik (1996, 1999, 2011), one of the strong factors contributing to referent activation is the distance to antecedent along the hierarchical discourse structure. One method of modeling the hierarchical discourse structure is Rhetorical Structure Theory (Mann & Thompson 1987, Taboada & Mann 2006). In accordance with this idea, the notion of rhetorical distance was proposed in Kibrik 1996, measuring the span between the current EDU and the antecedent EDU. Non-linear, hierarchical rhetorical structure is an approximation to cognitive structure, so rhetorical distance is a natural measurement reflecting the relationships between different states of the dynamically developing working memory.

Figure 1 contains the "filters" component. The most important of these filters is the speaker's check for potential referential conflict, that is, ambiguity of a projected referential expression anticipated by the speaker. As a result of this filter, a reduced referential expression may be ruled out even in the situation of high referent activation. We do not specifically focus on referential conflict at the present stage of our project.

3. The RefRhet2 corpus

We explore the multi-factorial referential choice with the help of an English-language corpus named RefRhet2. This corpus is based on the RST Discourse Treebank (Carlson, Marcu & Okurowski, 2003), selected because it was already annotated for rhetorical structure. Rhetorical annotation is highly labor-intensive, so it was very advantageous to make use of the already accomplished RST Discourse Treebank.

RefRhet2 currently contains 11,461 markables (=referential expressions), 3994 anaphor-antecedent pairs, 1115 third-person pronouns, 1696 descriptive noun phrases, and 1458 proper names. In RefRhet2, we implemented the MoRA (Moscow Referential Annotation) scheme, building upon the earlier PoCoS scheme (see Krasavina & Chiarcos, 2007). MoRA involves annotation of (i) markables, (ii) coreference, and (iii) features serving as potential activation factors. About 20 features, derived from the annotation or automatically computed, serve as candidate activation factors. They belong to four distinct groups, including referent's features (animacy, gender, person, number, protagonism), anaphor's features and antecedent's features (phrase type, syntactic role, etc.), and distances from an anaphor to its antecedent (in words, markables, elementary discourse units – linear and rhetorical, etc.).

One of the advantages of the MoRA scheme, as compared to some other annotation schemes (such as MUC-6 – Coreference Task Definition, 1995, GNOME – Poesio, 2000, and PoCoS – Krasavina & Chiarcos, 2007) is the annotation of groups, that is sets of markables that, collectively, serve as an antecedent of an anaphor. Group antecedents are a highly frequent phenomenon in RefRhet2, so an adequate account of this phenomenon is really crucial. We distinguish between coordinate, prepositional, and discontinuous groups and posit that both groups and their members can participate in referential chains. Figure 2 illustrates the annotation of a discontinuous group as an antecedent in the MMAX2 program.

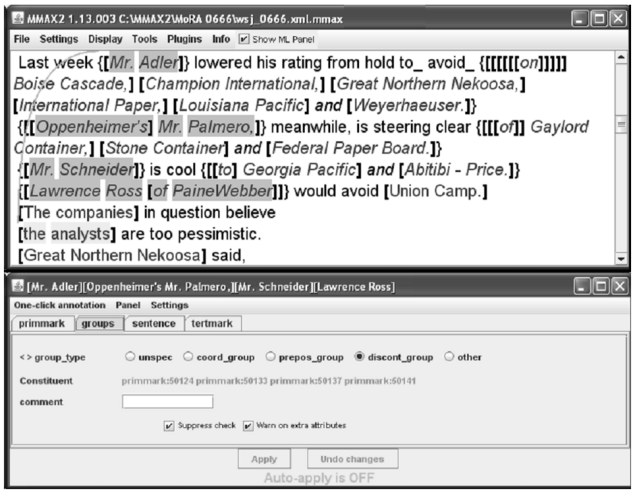

Figure 2: Annotation of a discontinuous group as an antecedent: Example from Text 0666

4. Computational modeling

We model referential choice with the help of machine learning algorithms. We use the WEKA system (Hall et al., 2009) that includes many algorithms, as well as automated means of algorithms' evaluation. Several types of algorithms have been employed in the present study: logical algorithms (decision trees C4.5, deciding rules algorithm JRip), logistic regression, and the so-called classifier compositions: bagging and boosting (Breiman, 1994; Freund & Shapire, 1996; Clark et al. eds. 2010).

In our computational model of referential choice the following two tasks were set. In the two-way task, we model the basic referential choice between a (third person) pronoun and a full noun phrase. In the tree-way task, we predict whether a given anaphor is a pronoun, a description, or a proper name. The criterion for choosing both an optimal set of features and an algorithm is accuracy, that is, the ratio of properly predicted referential expressions to the overall amount of referential expressions. The results of our modeling studies are shown in Table 1.

Table 1: Prediction of referential choice in RefRhet2: two-way and three-way tasks

Algorithm	Accuracy two-way	Accuracy three-way
Baseline (frequency of the most common referential option)	72.1%	37.2%
Deciding rules algorithm	85.9%	75.7%
Decision tree algorithm	87.1%	76.8%
Logistic regression	87.9%	77.6%
Bagging	88%	78.6%
Boosting	88.7%	78.7%

The best results were obtained from the boosting composition algorithm, providing the 88.7% accuracy for the two-way task and the 78.7% accuracy for the three-way task. We take these high results as an indication of the fact that our multi-factorial approach to referential choice is on the right track. Note that the high results in the three-way task were attained on the basis of the set of features originally developed for basic referential choice (two-way task).

The results of this study are based on a very laborious annotation process, involving many stages and many work hours. Is it possible to attain comparable results without fine-grained annotation? We have experimented with a cheap set of features that are either automatically extracted from texts or are in principle extractable. (An example of an annotated but extractable feature is the type of phrase, that is noun phrase vs. prepositional phrase; of course, prepositions can be listed exhaustively.) Working with the cheap set of features, we found that, in the case of logistic regression, the drop in accuracy was 2.7% for the two-way task and 2.1% for the three-way task. This suggests that a high level of accuracy can be attained within our multi-factorial approach even when modeling reference production in minimally annotated texts. Of course, annotation cannot be avoided completely; in particular, the annotation of markables and coreference is necessary in order to predict referential choice with high accuracy.

5. Non-categorical referential choice

In our study, we have achieved the accuracy of prediction, approaching 90% in the case of basic referential choice. A natural question arises: If we continue to refine our methodology, is it possible to further improve the results, bringing accuracy close to 100%? Most likely, the answer is negative. It is very often that referential choice is categorical, that is, one cannot replace a pronoun with a full NP or vice versa without rendering discourse infelicitous. However, in natural discourse there is usually a subset of instances in which two or more referential expressions are equally appropriate. Kibrik (1999), exploring an English narrative text, conducted an experiment in which certain referential expressions were altered and native speakers' judgments were used to find out which modifications were appropriate and which turned out infelicitous. A number of referential expressions were found to be alterable, and all of them corresponded to an intermediate level of referent activation. If that is correct, an algorithm cannot possibly predict referential choice with 100% accuracy. Suppose that our corpus contains 20% of alterable referential devices. Within the basic referential choice, that means that each of these referential devices can equally well appear as a pronoun or a full NP. Assuming that pronouns and full NPs occur with equal frequency, even an ideal algorithm must have 10% errors compared to the original referential choices. In fact, these instances should not count as errors; rather they are divergences from the original referential choices, but appropriate ones.

We believe that referential choice is less than fully categorical. Developing the approaches discussed in prior studies (such as Kibrik 1999, Belz & Varges 2007, van Deemter et al. 2012: 173–179), we have conducted an experiment addressing the instances of non-categorical referential choice. Nine texts were selected from the corpus, such that an original text contained a proper name, but a pronoun was predicted by the algorithms. All the 18 texts (including the original and the altered ones) were put on two experimental lists, so that each list contained only one version of the same text and the original and altered texts alternated. 60 participants answered questions about the referent encoded by a pronoun or a proper name. Questions to proper names were answered by the participants correctly 84% of the time. In seven out of nine texts, questions to pronouns were answered with the comparable level of correctness (80%), which demonstrates that pronouns were indeed appropriate in those discourse contexts.

We suggest that these instances correspond to the situations of intermediate referent activation (Kibrik 1999), making both a full NP and a pronoun possible. This hypothesis can be evaluated with the help of logistic regression that provides degrees of certainty of prediction. In our data, in all of the instances but one the degree of certainty in the prediction of a pronoun varied between 0.5 and 0.8. This interval can be interpreted as a moderate probability of a pronoun, according to the algorithm's judgment, and also as a correlate to a moderate referent activation level creating prerequisites for non-categorical referential choice.

6. Conclusions

Thus study addresses the question of the foundations of referential choice: how speakers select an appropriate device to mention a referent in discourse. We propose that the basic referential choice between full and reduced referential devices immediately depends on referent activation in working memory, and the latter is determined by multiple activation factors. We employ an annotated corpus from which many potential activation factors can be extracted. Algorithms of machine learning, supposedly simulating human behavior, predict referential choice with high accuracy. They perform well not only for the basic referential choice, but also for the three-way choice between

pronouns, proper names, and descriptions. However, it is not always the case that actual referential choice can be predicted with full precision. In natural discourse, there are instances in which more than one kind of referential device is appropriate. Non-categorical referential choice occurs in the situations of intermediate referent activation. We have conducted an experiment in which divergences between the human and computer's choices were subject to further human evaluation. The results of the experiment demonstrate that in the instances of divergences a referential device different from the original one is indeed appropriate. The degrees of prediction certainty provided by logistic regression lend further support to our analysis: in most of the instances of divergence the algorithm demonstrated a moderate level of certainty. Our approach, including both the multi-factorial prediction and the account of non-categorical choice, can be used in other natural language generation tasks.

Acknowledgements

We are grateful to Olga V. Fedorova and Natalia V. Loukachevitch who provided useful and important input during the preparation of this paper. The study was conducted with partial financial support from grant #11-04-00153 from the Russian Foundation for the Humanities.

References

Awh, E., Vogel, E.K., & Oh, S.-H. (2006). Interactions between attention and working memory. *Neuroscience, 139* , 201–208.

Belz, A., & Varges, S. (2007). Generation of Repeated References to Discourse Entities. In Busemann, S. (Ed.) *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG'07)* (pp. 9-16). Schloss Dagstuhl, Germany.

Breiman, L. (1994). *Bagging Predictors* (Technical Report 421), Department of Statistics, University of California at Berkeley.

Carlson, L., Marcu, D., & Okurowski, M.E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. van Kuppevelt & R. Smith (Eds.), *Current directions in discourse and dialogue*. Dordrecht: Kluwer.

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

Clark, A., Fox, C., & Lappin, S. (Eds.). (2010). *The handbook of computational linguistics and natural language processing*. Chichester: Wiley.

Coreference task definition, v2.3. (1995). *Proceedings of the Sixth Message Understanding Conference* (MUC-6), 335-344.

Fox, B. A. (1987). *Discourse structure and anaphora in written and conversational English*. Cambridge: Cambridge University Press.

Givón, T. (1983). Topic continuity in discourse: An introduction. In T. Givón (Ed.), *Topic continuity in discourse: A quantitative cross-language study*, (pp. 1-42). Amsterdam: Benjamins.

Freund, Y., & Schapire, R. (1996) Experiments with a New Boosting Algorithm. In L. Saitta (Ed.), *Machine Learning: Proceedings of the Thirteenth International Conference (ICML '96)*. Morgan Kaufmann.

Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P.,& Witten, I.H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations, 11*(1), 10-18. New York: ACM.

Kibrik, A. A. (1996). Anaphora in Russian narrative discourse: A cognitive calculative account. In B. A. Fox (Ed.), *Studies in anaphora* (pp. 255 – 304). Amsterdam: Benjamins.

Kibrik, A. A. (1999). Cognitive inferences from discourse observations: Reference and working memory. In K. van Hoek, A. A. Kibrik, & L. Noordman (Eds.), *Discourse studies in cognitive linguistics. Proceedings of the 5th International Cognitive Linguistics Conference* (pp. 29-52. Amsterdam: Benjamins.

Kibrik, A.A. (2011). *Reference in discourse*. Oxford: Oxford University Press.

Krasavina, O.N. & Chiarcos, Ch. (2007). PoCoS – Potsdam Coreference Scheme. In *Proceedings of the Conference of the Association for Computational Linguistics (LAW)* (pp. 156-163). Prague, Czech Republic.

Mann, W. C. & Thompson, S.A. (1987). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text 8*(3), 243-281.

Poesio, M. (2000) Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. In Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. & Stainhaouer G. (Eds.) *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May- 2 June, 2000.

Taboada, M., & Mann, W.C. (2006). Rhetorical structure theory : Looking back and moving ahead. *Discourse Studies, 8*, 423–459 .

van Deemter, K., A. Gatt, R. van Gompel & E. Krahmer (2012). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science* 4 (2), 166-183.