Possessor NPs and referential choice in English business prose (a corpus research)

Mariya Khudyakova

Moscow State University

The choice of an appropriate referential expression (definite description, proper name or pronoun) depends on multiple factors. This paper focuses on how the possessor position of a referential expression and its antecedent affect referential choice. other factors, such as syntactical role, form and definiteness of the antecedent, and animacy of the referent are considered. The study is based on a subcorpus of the specially designed RefRhet corpus.

Key words: corpus research, referential choice, anaphora, possessive pronouns

La elección de una expresión referencial conveniente (descripción definitiva, nombre propio o pronombre) depende de varios factores. En esta papel se aclara el problema de cómo la posición del poseedor de una expresión referencial y su antecedente influye en la elección referencial. Se examina también otros factores como el papel sintáctico, forma y definitividad del antecedente, el referente animado o no animado. La investigación se funda en un subcorpus del corpus especial RefRhet.

Palabras claves: investigación de corpus, elección referencial, anáfora, pronombres posesivos

1. Introduction

While producing discourse, a speaker constantly decides what referential expression to use to name a referent. Choosing an appropriate form of a language expression – full descriptive NP, proper name, pronoun, etc.—to refer to an object, person or abstract entity – is called referential choice. Referential choice is a complex cognitive process. For decades linguists have been making different models of referential choice, involving different syntactic, semantic, pragmatic and other factors.

Those factors can be divided into five groups: properties of the antecedent[39], of the referential expression and the referent, the nature of the relation between the antecedent and the anaphor, the genre of the text. Among the properties of the referential expression and its antecedent there are such factors as the syntactical role of the expression (Arnold, 2008; Kibrik, 2003), its semantic role (Rose, 2007), its phrase type (Kibrik, 1997). The properties of the referent, that can affect referential choice, are the animacy of the referent (Greenbacker & McKoy, 2009; Dahl & Fraurud, 1996) and the semantic properties of the referent (e.g. sortal classes in (Strube & Wolters, 2000). The relation between the referential expression and its antecedent is usually expressed in terms of distance, e.g. linear distance in sentences (Greenbacker & McKoy, 2009) or paragraphs (Kibrik, 1997), or rhetorical distance (Kibrik, 1997).

This paper focuses mainly on the role of a specific syntactical position of the referential expression—the possessor position. The two major questions are: 1) How does the referential choice (between full NPs and pronouns)in the possessor position happen? 2) Does the possessor position of the antecedent influence the referential choice of the anaphor?

A practical meaning of these tasks can be illustrated by the example. The language expressions 1, 2, 3 and 4 refer to the same object, that is, are coreferent.

> Tony pulls a tape measure across the front [of what was once a stately Victorian home]1. A deep trench now runs along [its]2 north wall, exposed when [the house]3 lurched two feet off [its]4 foundation during last week's earthquake.

The questions are the following: 1) Is there any difference in the possibility of pronominalization of the referential expressions in the possessor position (2 and 4) and non-possessor position (3)? 2) Does the possessor position of the NP2 affect the referential choice for the NP3?

2. Terminology

There is a certain inconsistence in the terminology dealing with possessor positions. The usual term for the pronouns referring to the possessor (*his*, *my*, etc) is "possessive pronouns". But also the term "possessive" is used for NPs referring to the possessor + object (*his car*, *John's house*, etc) (Willemse, 2009; Storto, 2007; Barker, 2000). In this paper the decision was made to name s-genitive and of-genitive full noun phrases and pronouns, which refer to the possessor, "possessor full NPs" and "possessor pronouns"

[39] Antecedent and anaphor are coreferent expressions, the antecedent being the closest one in the previous context to the anaphor.

respectively. The non-possessor pronouns and full NPs are called actant pronouns and actant full NPs.

3. *RefRhet corpus*

The research is based on the specially annotated RefRhet corpus which consists of 385 Wall Street Journal articles (Kibrik, Dobrov, Zalmanov, Linnik and Loukachevitch, 2010) The RefRhet corpus is based on the English-language corpus RST Discourse Treebank, created under the direction of Daniel Marcu (http://www.isi.edu/~marcu/discourse/Corpora.html), see (Carlson, Marcu, Okurowski and 2003). The corpus contains 176 383 words.

Referential annotation was added to RST Discourse Treebank, and as a result the RefRhet corpus emerged. Referential annotation was performed with the help of a so-called annotation scheme, see (Krasavina & Chiarcos, 2007). The annotation scheme employed contains a set of annotated parameters, or factors.

An element that undergoes annotation, called markable, is a text constituent that can serve as a referential expression. Coreference relations are posited between markables. In addition, each markable contains a number of annotated features (grammatical role, animacy, etc.) that can affect referential choice.

Since all of the annotations are performed manually, a certain number of mistakes is inevitable. In order to exclude such mistakes the decision has been made to annotate each text twice and then compare these annotations automatically. Such comparison results in a list of markables that either appear only in one of the annotations, or have different feature values in the two annotations. Subsequently, annotators from a different group choose the correct analysis out of the two available.

The present-day stage of the RefRhet corpus is as follows: 157 texts are annotated twice, 193 texts are annotated once, and 25 texts are not yet annotated.

For the research a subcorpus of 31 text was chosen. These are the texts that had been annotated twice, and also the procedure of the comparison and correction of the annotations was performed. The subcorpus contains 3453 markables. Since the current annotation scheme suggests the annotation of possessor pronouns, but not of possessor full NPs, the cases of s-genitive and of-genitive were annotated in the subcorpus. In order to exclude the cases of the reflexive possessor pronouns, that are rather syntactical than the result of the referential choice (Bach & Partee, 1980), the possessor pronouns, whose antecedents were in the same clause, were not taken into consideration.

The correlation between different factors and the referential choice was elicited with the help of log-linear analysis, which is used for establishing the correlation of two or more factors.

4. *Results*

There are 3092 definite NPs in the subcorpus of RefRhet, 85% of which are full NPs, and only 15% are pronouns. NPs in the possessor position present 19% of the chosen markables. The distribution of the types of referential expressions in possessor and actant position are in Table 1.

Table 1. The distribution of full NPs and pronouns in the possessor and actant position in the subcorpus.

	Possessor position		Actant position		total	
Full NPs	259	8%	2253	73%	2512	81%
pronouns	213	7%	367	12%	580	19%
total	472	15%	2620	84%	3092	100%

As can be seen from Table 1, possessors are more likely to be pronominalized than actants.

Also there is a strong correlation between the animacy of the referent, referential choice and the possessorness of the referential expression. Animate possessors are pronominalized in 80% cases while inanimate possessors and actant NPs are more likely to be full NPs. The most predictable is the referential choice for inanimate referents in actant positions.

The most interesting cases are when possessors and actants demonstrate contrary tendencies to be pronominalized depending on some properties of the antecedents, for example, the definiteness of the antecedent. As can be seen in Table 2, possessors are more likely to be expressed as pronouns after indefinite NPs, and after definite antecedent the numbers of possessor pronouns and full NPs are almost equal, while pronominalized actants have a contrary tendency: they are more likely to be full after definite antecedents.

Table 2. the distribution of full NPs and pronouns in the possessor and actant position in the subcorpus and the definiteness of their antecedents.

antecedent	anaphor											
	possessor						actant					
	Full NPs		Pronouns		**total**		Full NPs		Pronouns		**total**	
Definite NPs	174	47%	195	53%	**369**	**100%**	912	73%	344	27%	**1256**	**100%**
Indefinite NPs	5	29%	12	71%	**17**	**100%**	73	60%	48	40%	**121**	**100%**
total	**386**						**1377**					

There is also a correlation between the form of the antecedent and referential choice in possessor / actant position. Possessors after full NPs are pronominalized less often than after pronouns. The pronoun form of the antecedent has a strong effect on the form of the possessor anaphor—such anaphors are pronouns in ¾ of the cases. Actants have a contrary tendency: they are more likely to be full NPs, than pronouns, especially after full NP antecedents.

The correlation between the syntactical role of the antecedent and the form and possessorness of the anaphor is also statistically significant. Actant NPs demonstrate a tendency to be full NPs after all types of antecedents, while possessor NPs are pronouns after 74% of subject antecedents, 52% of direct object antecedents and 40% of other antecedents.

There is no significant correlation between the possessorness of the antecedent and referential choice.

The research has shown that such factor as possessor / actant position of the antecedent and the anaphor affect referential choice. This feature will be added to the annotation scheme of RefRhet.

References

Arnold, J. (2008). Reference Production: Production-internal and Addressee-oriented Processes. Language and Cognitive Processes. Retrieved from http://www.unc.edu/~jarnold/pages/publications.html

Bach, E. & Partee, B. (1980). Anaphora and semantic structure. In B. Partee (ed.), *Compositionality in Formal Semantics - Selected Papers by Barbara H. Partee*. Blackwell.

Barker, Ch. 2000. Definite Possessives and Discourse Novelty. In *Theoretical Linguistics Volume 26 (3)*. De Gruyter

Carlson, L., Marcu, D. and Okurowski, M. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt and R. Smith (eds.) *Current directions in discourse and dialogue*. Dordrecht: Kluwer, 2003. Pp. 85–112.

Chiarcos Ch. & Krasavina O. (2005). Annotation Guidelines. PoCoS — Potsdam Conference Scheme. Draft.

Dahl, Ö. & Fraurud, K. (1996). Animacy in Grammar and Discourse. In Th. Fretheim & J. Gundel (eds.) *Reference and Referent Accessibility*. Amsterdam/Philadelphia: John Benjamins.

Greenbacker, Ch. & McCoy, K. (2009). Feature Selection for Reference Generation as Informed by Psycholinguistic Research. In *Proceedings of the 2009 Workshop on Production of Referring Expressions (PRE-CogSci 2009)*, Amsterdam.

Kibrik, A. (1997). Modelirovanie mnogofaktornogo protsessa: model referentsialnogo sredstva v russkom yazyke. *Vestnik MGU, 1997.4*, 94-105.

Kibrik, A. (2003) Analiz diskursa v kognitivnoy perspective. PhD thesis.

Kibrik, A., Dobrov, G., Zalmanov, D., Linnik, A. and Loukachevitch, N. (2010). Referencial'nyj vybor kak mnogofaktornyj verojatnostnyj process [Referential choice as a multi-factor probabilistic process]. In Aleksandr E. Kibrik (ed.), *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2010)*. Bekasovo, Moscow region. Moscow: RGGU, 173–181.

Rose, R. (2007). Pronoun Resolution and The Influence of Syntactic and Semantic Information on Discourse Prominence. In Branco, A. (ed.), *Anaphora: Analysis, Algorithms and Applications* (pp. 28-43). Berlin: Springer-Verlag.

Storto, G. (2007). On the structure of indefinite possessives. In B. Jackson and T. Matthews (eds.), *Proceedings of Semantics and Linguistics Theory X*. Ithaca, NY: CLC. 2007.

Strube, M. & Wolters, M. (2000). A Probabilistic Genre-Independent Model of Pronominalization. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics, April 29-May 04, 2000, Seattle, Washington.*

Willemse, P. (2007). Direct and indirect anaphora and the possessee referent of possessive NPs in English. In A. Branco, T. McEnery, R. Mitkov and F. Silva (eds.) *Proceedings of DAARC 2007 (6th Anaphora and Anaphora Resolution Colloquium)*. Porto: CLUP.