Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2017”

Moscow, May 31—June 3, 2017

SEMANTIC CLASSIFICATION OF RUSSIAN HETERONOMINATIVE NOUN PHRASES ON THE MATERIAL OF RUCOR CORPUS

Azerkovich I. L. (ilazerkovich@edu.hse.ru)

National Research University Higher School of Economics,
Moscow, Russia

Mentions of a referent usually vary over the span of text, and sometimes an NP, entirely different from it can be used. Automated coreference analyzers may not join such varying nominations of a referent within text, leading to loss in quality. This problem can be solved by using semantic information for analysis. This article is written as a preparative step of a research, dedicated to integrating semantic information in algorithms of automated coreference analysis for Russian. It presents a typology of varying nominations of a referent across text, based on type of their relation to the head of the chain. It differentiates between mentions that require encyclopedic or situational knowledge, and discusses sources of data that can help to properly process such examples. The obtained results can help in employing information from external sources for coreference resolution by allowing to evaluate a potential increase in quality and to understand what kind of information should be used in certain cases.

Keywords: natural language processing, coreference resolution, corpus analysis, semantic information

СЕМАНТИЧЕСКАЯ КЛАССИФИКАЦИЯ ГЕТЕРОНОМИНАТИВНЫХ ИМЕННЫХ ГРУПП В РУССКОМ ЯЗЫКЕ НА МАТЕРИАЛЕ КОРПУСА RUCOR

Азеркович И. Л. (ilazerkovich@edu.hse.ru)

Национальный исследовательский институт
«Высшая школа экономики», Москва, Россия

1. Introduction

Coreference resolution is a very important step of natural language processing, which involves joining together in a single *coreference chain* all mentions of a referent (*head of the chain*). Realization of this task, though, causes certain difficulties. One such example that often gets the attention of researchers, are recall errors, caused by automated analysis systems processing mentions, not identical to the head, as not coreferent to it. For example, during the RU-EVAL-2014 evaluation campaign [Toldova et al 2014], coreferent pairs such as *Grigori Perelman—matematik* 'a mathematician' or *Grigori Perelman—rossijanin* 'a Russian' presented difficulty for analyzers.

In theoretical linguistics, the problem of referential choice and factors influencing it are widely discussed (e.g. [Prince 1981], [Chafe 1994], [Kibrik 2001]). The consensus, which all researchers agree upon, is that depending on activation, or degree of attention focused on a referent, one way to mention it or another is chosen. The main distinction is drawn between descriptive NPs and anaphoric expressions, but variation within descriptions is not always considered. The question this discussion presents for applied linguistics is how to join all mentions of an entity in a single coreferent chain.

On Russian material, the problem of different NPs describing the same entity was addressed, for example, by N. Arutyunova. In her paper [Arutyunova 1987] she suggests referring to such expressions as *"heteronominative"*, and consequently this phenomenon itself can be called *"heteronomination"*. This terminology will be adopted further in this article. Arutyunova singles out causes of heteronomination occurring, such as speech usage, pragmatical factor, and figurative descriptions. She also suggests classifying heteronominative NPs by their function into introductory (existential), identifying (referential), predicative and vocative.

A detailed description of rules of nomination construction was presented in the thesis of [Toldova 1994]. Among others, the paper suggests drawing a distinction between thesaurus and non-thesaurus information, or between the set of classes an object belongs to, and information about it, drawn from the frame it is situated in, or from the current situation itself. A distinction is also drawn between information obtained from databases on the one hand, and information obtained metatextually or situationally on the other, which can be correlated to different types of data used for coreference resolution, as discussed below.

Another typology of coreferent NPs is presented in the article [Recasens et al 2010]. In this work, the notion of near identity is introduced to refer to NPs that are included in the same coreference chain but refer to different discourse entities. It means that meaning of an NP in the discourse used to refer to the head of the chain, is different from the head's one. In the article a classification of identity classes of mentions is presented: non-identity, identity and near-identity, with cases of near identity being classified, based on type of relation to the chain head.

To effectively join such cases of nominationstogether, various types of semantic information can be employed. This is demonstrated in e.g. [Harabagiu et al 2001] or [Rahman and Ng 2011]. Nevertheless, not many researchers consider improvement in quality, gained by resolving coreference for such cases, when working with Russian material (one of a few examples is presented in [Bogdanov et al, 2014], which describes a commercial system using an existing ontology).

As a preparative step in a research, dedicated to integrating semantic information in automated coreference resolving systems for Russian, it was necessary to distinguish different types of relations between NPs in a coreferent chain and their heads. This article describes such classification of heteronominative NPs in Russian texts, based on type of their semantic or pragmatic relation to their heads. To achieve this goal, NPs extracted from the annotated corpus of Russian texts were analyzed. The achieved results allowed to, firstly, evaluate a possible increase in quality from considering this data in analysis, and, secondly, to understand relative frequency of different relation types between mentions of the same referent. This data can give a better understanding of effective usage of semantic information for automated analysis.

2. Research material

As the source of texts for analysis, Russian coreference corpus RuCor was used. This corpus was created for the task of automated anaphora and coreference resolution for Russian RU-EVAL-2014. It contains 185 texts: mostly news articles, but also fiction, blog posts and Wikipedia articles, consisting of 3,633 coreferent chains and 16557 coreferent groups in total. [Toldova et al 2016] It also includes annotation of coreferent chains, done by hand as the Gold Standard for the task, so this information was precise enough to be used for my research.

For each annotated corpus chain its type is noted, either *anaphoric* (containing only the head and its anaphor), or *coreferent* (containing several NPs). For each element in the chain its POS-tags and syntactic position are annotated: *predicative*, *appositive*, *direct speech* or *definitive* (other).

The first step in material preparation was selecting all coreferent chains from the corpus. Then in each chain NPs that did not coincide with the head were selected, as their coreference could only be established by using semantic information, and not string-based features. Finally, for each of the selected groups type of its relation to the head was annotated: was textual or encyclopedic information required to establish coreference, and the exact type of semantic relation.

Of 4,435 coreferent chains, currently present in the corpus, 798 included heteronominative NPs (see Fig. 1). Number of NPs that were classified as heteronominative was 1,695, or approximately 11% of total count of NPs in the corpus. Moreover, among two-element coreferent chains, the second element of ⅓ of them was heteronominative. This shows that discarding semantic information in task of coreference resolution we risk a serious recall loss, both in chain count and length.

Increase of number of heteronominations in a chain with its length was also observed. The tendency can be seen, though with fluctuations due to number of occurrences within corpus, on chains with lengths from 2 to 12, comprising 79% of it (Fig. 2). Correlation coefficient between the two values equals 0.9, which represents a strong connection. This dependency can be explained by the fact that a longer coreferent chain makes possible usage of incomplete descriptions, bringing forth more aspects of the referent.

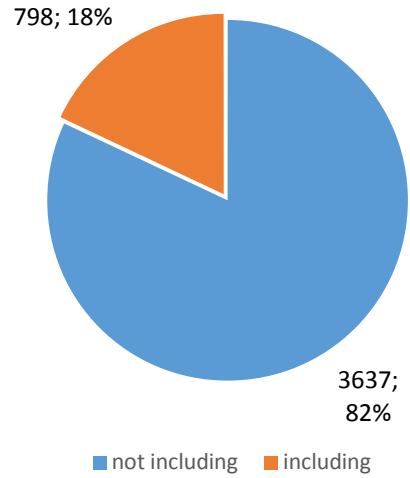

Fig. 1. Number of coreferent chains including and not including heteronominative NPs

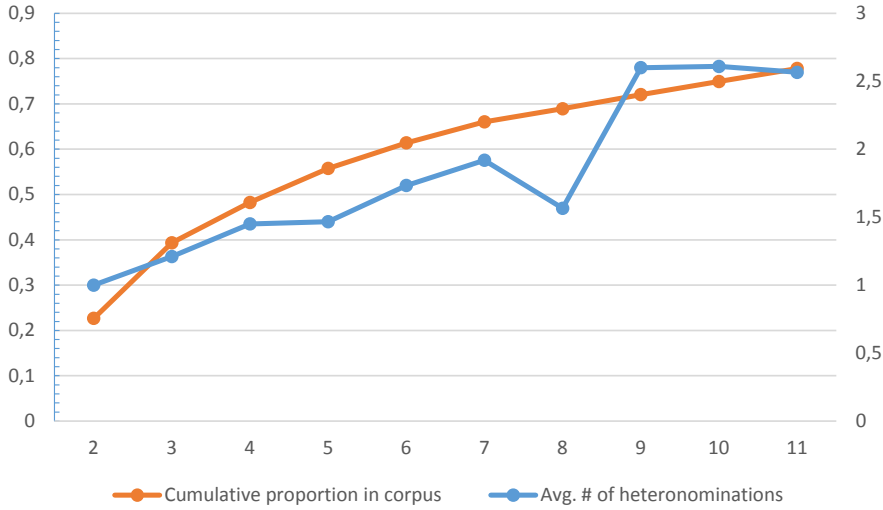

Fig. 2. Average number of heteronominations and cumulative proportion of coreferent chains in corpus by length

3. Classification

In this article, a classification of coreference cases is presented, for proper resolution of which it would be effective to employ semantic information, or, precisely, the type of semantic or pragmatic relation between the NP and the head of its coreferent chain. Because this classification was created with practical use of ontologies and other sources of encyclopedic knowledge for coreference resolution in mind, it was based on straightforward relations between two noun phrases that occurred in text. In this regard, it differs from typologies described in the Introduction above. Despite similar problematics, it does not fully correspond to cases of near identity described in [Recasens et al 2010], because in the article an accent is made on discursive relations between the two NPs. The more suitable classification, on which the final version presented here is partly based, was described in [Toldova 1994], where an opposition of encyclopedic vs. situational and metatextual information was suggested.

Based on this opposition, the NPs from the corpus were first classified by the general type of information, required to establish a coreference relation to their heads. Two such types were singled out: *textual*, or pragmatic, information and *encyclopedic*, or ontological, information. The second class of NPs was then further divided into sub-classes, depending on the exact nature of the relation between two NPs.

Encyclopedic information here means presence or absence of an ontological relation between the NP and the head of its coreferent chain. To discover such a connection, external sources of knowledge should be consulted. The most popular ones are Wikipedia and WordNet, because both contain information about a large amount of entities, are periodically renewed, and possess explicit structure, which allows for easy extraction of hierarchic relations. Technology of using these data is well-developed and described in many works, e.g. [Bunescu and Pasca 2006] or [Strube and Ponzetto 2006].

Textual information, on the other hand, is required to determine coreference relations of NPs, not ontologically connected to the head of a chain. In terms used in [Toldova 1994] it can be classified as covering the types of non-thesaurus situational and metatextual (anaphoric or demonstrative) information. Such NPs in many cases carry a positive or negative sentiment or express the relation of the referent to the situation described in text. Consequently, they rely heavily on the preceding context to establish proper relation to their referent, either repeating NPs already used for description or alluding to a part of the preceding narrative. For example, in (1) coreference between *Agapovoj* and *nashej geroinej 'our heroine'* can be inferred from the fact that a character from fiction is its hero.

(1) *Poprobuju* [*jejo*]$_i$ *izobrazit'. Hotja vneshnost'* [*Agapovoj*]$_i$ *sushchestvennogo znachenija ne imeet. … Sledujem za* [*nashej geroinej*]$_i$.
I'll try to picture [*her*]$_i$. [*Agapova's*]$_i$ *appearance has no special meaning, though. … We follow* [*our heroine*]$_i$.'

Because the relation between two NPs is not encyclopedic, no external sources need to or can be consulted. To determine coreference in such cases, information obtained from the text itself, such as appositive constructions or word co-occurrences, can be used ([Rahman and Ng 2011], [Bansal and Klein 2012], et al). Advantages

of this kind of data are not relying on information from additional sources, and obtainability from unannotated texts, which allows to save time on pre-processing.

The resulting classification which was obtained as the result of my research consists of the following six types of relations: hypo/hypernymic, synonymic (with a separate class of diminutive forms), cohyponymic, metonymic and textual. Their relative distribution in texts of the corpus is presented in Fig. 3.

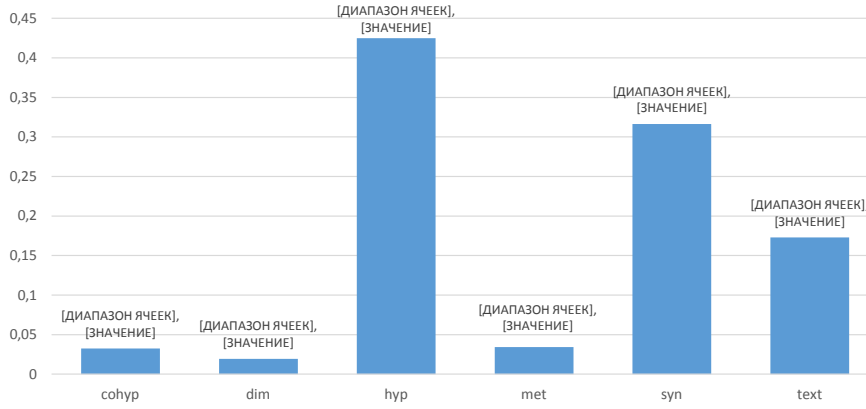

Fig. 3. Distribution of relation types in texts

3.1. Hypo/hypernymic relation

As can be seen in Fig. 3, semantic relations of this type are the most represented in the corpus and constitute almost a half (42%) of all cases of heteronomination. Most often it joins the NPs that either introduce a referent in the discourse (existential nominations), or act as a part of an identifying nomination. In journalistic or news texts this class is mostly presented by characteristics of named entities that either have been previously named in the text, or are unique at any given moment in time. Posts and positions were attributed as hyponyms instead of synonyms, because they are not inherent to the person they refer to in text, and their referent can change with time. For example, in (2) the sentence starts with the mention of a person's position, and the current Mayor of Moscow is Sergei Sobyanin. But, because when the text was published this position belonged to Yuri Luzhkov, the referent of the NP can be correctly deduced.

(2) *Protivostojanije* [*stolichnogo gradonachal'nika*]$_i$ *i Federal'nogo Orgkomiteta po podgotovke prazdnovanija 65-letija Pobedy zakonchilos' absolutnoj pobedoj* [*Yurija Luzhkova*]$_i$.
'The conflict between [*the Mayor of the capital*]$_i$ *and the Federal committee on preparation for celebration of 65-th anniversary of the Victory ended with* [*Yuri Luzhkov's*]$_i$ *absolute victory.'*

This relation corresponds most closely to "thesaurus information" in [Toldova, 1994], as it is described precisely as "a system of generalized frames, corresponding

to classes of objects". In comparison, in [Recasens et al, 2010] such relations are described as "name metonymy", because reference is transferred from an object to one of its facets.

3.2. Synonymic relation

This type of relation was the second most frequent in the corpus, and together with hyponimic relation described above, they constitute almost ¾ of all examples (74%). It describes relations between synonyms, or words that would have a common parent node in an ontology applied for coreferential analysis. Abbreviations of existing entities were also attributed to this type, for example: *Russian Federation—Russia—RF 'Russian Federation'*. In [Recasens et al, 2010] synonymy is considered a case of full identity, but, as evidently heteronominative, such NPs were included in my analysis.

(3) *V ramkah festivalja, posvjashchennogo* [*legendarnomu spektaklju "Dobryj chelovek iz Sezuana"*]$_i$*, teatr pokazal* [*znamenituju postanovku*]$_i$ *v dvuh cheshskih teatrah.*
'During the festival, devoted to [*the legendary play "The Good Person of Szechwan"*]$_i$*, the theatre played* [*the famous piece*]$_i$ *in two Czech theatres.'*

3.3. Diminutives

Diminutive NPs in the corpus were only encountered in literary texts. They were classified as a separate relation type from synonyms, because they are closer in form to the head of coreferential chain. Being derivatives, words of these group only differ from their respective heads due to the presence of a suffix. Because of this, it might be possible to infer coreference between a diminutive and its head without consulting encyclopedic sources and only using string features, such as token similarity. While diminutives do not constitute a large group compared to other relation types, processing time for the analysis can be saved if no semantic information is used.

(4) [*Van'ka*]$_i$ *s* [*Tan'koj*]$_j$*, tochnee skazat',* [*Ivan Tihonovich*]$_i$ *i* [*Tat'jana Finogenovna*]$_j$ *Zaplatiny, vecherami ljubili posidet' na skamejke vozle svojego doma.*
'[*Van'ka*]$_i$ *and* [*Tan'ka*]$_j$*, or rather* [*Ivan Tihonovich*]$_i$ *and* [*Tat'jana Finogenovna*]$_j$ *Zaplatiny, loved to sit on the bench near their house in the evenings.'*

3.4. Metonymic relation

This type of relation to the head of coreferential chain is characteristic for NPs, referring to an element, part or material of the head. News texts especially abound in examples of metonymy (80% of texts, containing metonymic groups belonged to this genre), with possible NPs for referring political entities varying from a governing body to the capital of a country to the whole country. At the same time, these steps are rarely present in a single text, and most occurrences are either of "body-country" type, e.g. *kitajskije vlasti 'Chinese government'—Kitaj 'China'*, or of "capital-country" type, e.g. *Suhum 'Sukhumi—Abhazija 'Abkhazia'* (5).

(5) *Nezavisimost' Suhuma priznala tol'ko RF* (*i chastichno Nikaragua, ne ustanoviv s Abhazijej diplomaticheskih otnoshenij I ne ratificirovav dekret prezidenta o priznanii cherez parlament*).
'Sukhumi's independency was acknowledged only by RF (*and partly Nicaragua, without establishing diplomatic relations with Abkhazia or ratifying president's decree about recognition through parliament*)'

It should be noted that in some cases of material-object metonymy the same approach as for diminutives (Section 3.3) can be applied. In pairs such as *brilliantovaja brosh' 'a diamond brooch'—eti brillianty 'these diamonds'* both elements are also related, which makes it possible to determine coreference using string features only.

3.5. Cohyponimic relation

This type of relations includes words that would have a common parent node in an ontology. For example, in dialogue from (6) both NPs *pozhilaja kolhoznitsa 'elderly farmer woman'* and *mamasha 'mother'* have a common parent node 'woman'.

(6) *— V takoj-to temnote ne ugljadish', — skazala* [*pozhilaja kolhoznitsa*]$_i$
— Hot' by skorej svet dali.
— Nichego, [*mamasha*]$_i$*, v tesnote da ne v obide.*
'— You can't see when it's so dark, — said [*the old farmer woman*]$_i$.
— I wish they turned on the light.
— Don't worry, [*mother*]$_i$*, the more the merrier.'*

The difference between cohyponymic and textual relation lies in the fact that while both types can be used to provide additional descriptions of referents, the former is more often used for non-referential NPs, such as *aktivisty 'activists'—protestujushchije 'protesters'*, or existing entities, e.g. *Barack Obama—superzvezda 'superstar'*, while the latter is more often used for characters of a single text.

Still, given that cohyponyms are rarely present in the corpus, it might be more profitable in terms of computation time and data amount to rely only on the texts for analysis.

3.6. Textual relation

As has already been mentioned, relation of this type occurs most often in fiction or journalistic texts. In texts of the corpus NPs that are joined by it usually provide additional characterization for personages of the narrative or rephrase the context in which the referent occurred previously.

The question is open to what extent information about literary characters can be considered textual rather than encyclopedic, but information that can be attributed as "situational", or related to the content of the text itself, certainly cannot be obtained from ontologies. E.g. in (5), *rodina 'homeland'* refers to Nigeria rather than Israel, only because it is stated previously in the sentence that the hero of the article was born there.

(7) *Vyhodec is* [*Nigerii*]$_i$ *reshil ostat'sja na PMZH v Izraile, poskol'ku na* [*rodine*]$_i$ *ego jakoby presledujet opasnyj prizrak.*
'A native of [*Nigeria*]$_i$ *decided to permanently reside in Israel, because in his* [*homeland*]$_i$ *he is allegedly followed by a dangerous ghost.'*

For NPs belonging to this class, predicative position is preferable more than for others. Despite this relation occurring in the corpus almost 2.4 times fewer than hyponymy, NPs of the former take the predicative position 2.5 times more often (12.97% vs 5.15%), and they also represent a slightly larger part of examples for this syntactic position in general (32.5% vs 31.6%, respectively). Unfortunately, statistics for NPs in appositive position were not immediately available, but this data inclines to suggest that NPs textually related to their head would be more frequent in this position, as well.

4. Conclusions

After conducting the described above research, the following results were obtained:

1. A classification of cases of heteronomination in Russian texts was built.
2. It was ascertained, that heteronominations occur in texts frequently enough for their omission to have considerable negative impact on quality of coreference resolution.
3. Dependency of amount of heteronominations in a coreference chain and its length was observed.
4. It was determined that encyclopedic information would be required for coreference resolution of heteronominative groups more often that textual.

This paper presented an important step in improving quality of automated coreference resolution for Russian by employing semantic knowledge in the process. Because the classification described here is partly based on relation types that are used in ontology building, it can be easily applied to obtain data from a source of encyclopedic information and use this data for automated coreference resolution. Among plans for future work, based on the result of this research, is also modifying existing algorithms for coreference resolution in Russian to accommodate semantic and pragmatic features, and possibly a creation of an open-source database, accumulating ontological information for possible uses in natural language processing of Russian.

References

1. *Arutjunova, N. D.* (1987), Nomination and Text [Nominatsia i text], Language Nomination [Yazikovaja Nominatsia], Nauka, Moscow, pp. 304–357.
2. *Bansal, M., Klein, D.* (2012), Coreference semantics from web features, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Association for Computational Linguistics pp. 389–398.

3. *Bogdanov A. V., Dzhumaev S. S., Skorinkin D. A., Starostin A. S.* (2014). Anaphora Analysis based on ABBYY Compreno Linguistic Technologies, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2014" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2014"], Bekasovo, pp. 89–102.
4. *Bunescu, R. C., Pasca, M.* (2006), Using Encyclopedic Knowledge for Named Entity Disambiguation, Eacl (Vol. 6), pp. 9–16.
5. *Chafe W.* (1994), Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing, University of Chicago Press.
6. *Harabagiu Sanda M, Razvan Bunescu, and Steven J Maiorano* (2001), Text and knowledge mining for coreference resolution, Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. Association for Computational Linguistics, pp. 1–8.
7. *Kibrik, A. A.* (2001), Reference maintenance in discourse, Language typology and language universals. An international handbook, 2, pp. 1123–1141.
8. *Ponzetto, S. P., Strube, M.* (2006), Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution, Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Association for Computational Linguistics, pp. 192–199.
9. *Prince E. F.* (1981), Toward a taxonomy of given-new information, Radical pragmatics, Academic Press, 1981.
10. *Rahman A., Ng V.* (2011), Coreference resolution with world knowledge, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, Association for Computational Linguistics pp. 814–824.
11. *Recasens, M., Hovy, E. H., Martí, M. A.* (2010). A Typology of Near-Identity Relations for Coreference (NIDENT), LREC.
12. *Toldova S. Ju.* (1994), Structure of discourse and mechanism of focusing as important factors of choice of object nomination in text [Struktura diskursa I mehanizm fokusirovanija kak vazhnie faktori vibora nominatsii ob'ekta v tekste], Candidate thesis.
13. *Toldova S. Ju., Roytberg A., Ladygina A. A., Vasilyeva M. D., Azerkovich I. L., Kurzukov M., Sim G., Gorshkov D. V., Ivanova A., Nedoluzhko A., Grishina Y.* (2014), RuEval-2014: Evaluating Anaphora and Coreference Resolution for Russian, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2014" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2014"], Bekasovo, pp. 681–694.
14. *Toldova S., Grishina Y., Ladygina A., Vasilyeva M., Sim G., Azerkovich A.* (2016), Russian Coreference Corpus, Input a Word, Analyze the World: Selected Approaches to Corpus Linguistics, Cambridge Scholars Publishing, pp. 107–124.