ИНТЕЛЛЕКТ ЯЗЫК КОМПЬЮТЕР

Выпуск 18
Том II

ТРУДЫ

МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ ПО КОМПЬЮТЕРНОЙ И КОГНИТИВНОЙ ЛИНГВИСТИКЕ

TEL–2018

Казань, 31 октября – 3 ноября 2018 г.

INTELLECT. LANGUAGE. COMPUTER

ISSUE 18

PROCEEDINGS OF COMPUTATIONAL MODELS IN LANGUAGE AND SPEECH WORKSHOP

(CMLS 2018)

CO-LOCATED WITH THE 15TH
TEL INTERNATIONAL CONFERENCE ON COMPUTATIONAL
AND COGNITIVE LINGUISTICS (TEL-2018)

Volume 2

KAZAN, RUSSIA
October 31 – November 3
2018

УДК 004.8+81’32
ББК 81.1

Академия наук Республики Татарстан
Институт прикладной семиотики

Казанский (Приволжский) федеральный университет
Российский фонд фундаментальных исследований

Издание осуществлено при финансовой поддержке
Российского фонда фундаментальных исследований совместно
с Правительством Республики Татарстан
(проект №1 8-47-161001)

Научные редакторы:
доктор физико-математических наук **А. М. Елизаров**,
доктор технических наук **Н. В. Лукашевич**

Т78 **Труды международной конференции по компьютерной и когнитивной лингвистике TEL-2018.** – В 2-х томах. Т 2. – Казань: Изд-во Академии наук РТ, 2018. – 203 с.

ISBN 978-5-9690-0478-8

Сборник содержит материалы международной конференции по компьютерной и когнитивной лингвистике TEL-2018 (Казань, 31 октября – 3 ноября 2018 г.).

Для научных работников, преподавателей, аспирантов и студентов, специализирующихся в области компьютерной и когнитивной лингвистики и ее приложений.

УДК 004.8+81’32
ББК 81.1

ISBN 978-5-9690-0478-8

Proceedings of the International Workshop on
Computational Models in Language and Speech (CMLS 2018)
co-located with the 15th International Conference on
Computational and Cognitive Linguistics (TEL-2018)

Originally published online by CEUR Workshop Proceedings (CEUR-WS.org, ISSN 1613-0073), vol.2303.

Preface

This volume contains the papers presented at International Workshop on Computational Models in Language and Speech co-located with the 15th International Conference on Computational and Cognitive Linguistics (TEL-2018, http://telconf.tatar). The TEL-2018 conference held from October 31 to November 3, 2018 in Kazan, Russia. The conference and the workshop are organized by the Institute of Applied Semiotics of Tatarstan Academy of Sciences and Kazan Federal University.

The goal of this workshop is to bring together leading researchers from artificial intelligence, computational linguists, software researchers that are interested in natural language processing – both as speakers and as audience members. Its ultimate goal is to share knowledge, discuss open research questions, and inspire new paths.

The scope of "Computational Models in Language and Speech" workshop includes the following topics: Semantic analysis of the text, Semantic Web technologies, Thesauri, ontologies, Machine translation, Natural Language Processing, Speech technologies.

We received 22 submissions describing new research for the workshop. Collected papers have undergone preliminary reviewing. The acceptance rate for full papers was 59% (11 full-papers and 2 short papers were accepted).

The TEL-2018 conference and this workshop were supported by the Russian Foundation for Basic Research, the project # 18-47-161001.

We would like to thank all who contributed to our workshop. First of all, we thank the authors for submitting their high-quality research works to the workshop. We would like to thank the members of the program committee for their valuable review contributions. We are grateful to our organizing committees, who made the conference possible.

Program Committee Chair

Alexander Elizarov**,** Dr.Sc., Kazan (Volga Region) Federal University, N.I. Lobachevsky Institute of Mathematics and Mechanics, Kazan, Russia.

Program Committee and Reviewers

1. Natalia Loukachevitch, Dr.Sc., Lomonosov Moscow State University, Research Computing Center, Moscow, Russia
2. Dmitry Lande, Dr.Sc., Institute for Information Recording, National Academy of Sciences of Ukraine, Kyiv, Ukraine
3. Olga Nevzorova, Cand.Sc., Tatarstan Academy of Sciences, Institute of Applied Semiotics, Kazan, Russia
4. Valery Solovyev, Dr.Sc., Kazan (Volga Region) Federal University, Leo Tolstoy Institute of Philology and Intercultural Communication, Kazan, Russia
5. Vladimir Polyakov, Cand.Sc., Russian Academy of Sciences, Institute of Linguistics, Moscow, Russia
6. Mikhail Kopotev, Ph.D., University of Helsinki, Department of Modern Languages, Helsinki, Finland
7. Airat Khasianov, Ph.D., Kazan (Volga Region) Federal University, Higher School of Information Technologies and Information Systems, Kazan, Russia
8. Dzhavdet Suleymanov, Dr.Sc., Tatarstan Academy of Sciences, Institute of Applied Semiotics, Kazan, Russia
9. Alexander Kirillovich, Kazan (Volga Region) Federal University, N.I. Lobachevsky Institute of Mathematics and Mechanics, Kazan, Russia
10. Yury Zagorulko, Cand.Sc., A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
11. Elena Tutubalina, Cand.Sc., Kazan (Volga Region) Federal University, Higher School of Information Technologies and Information Systems, Kazan, Russia
12. Aidar Khusainov, Cand.Sc., Tatarstan Academy of Sciences, Institute of Applied Semiotics, Kazan, Russia
13. Alexander Fridman, Dr.Sc., Institute for Informatics and Mathematical Modeling – Subdivision of the Federal Research Centre «Kola Science Centre of the Russian Academy of Sciences»
14. Sergei Tatevosov, Dr.Sc., Lomonosov Moscow State University, Faculty of Philology, Moscow, Russia

Organizing Committee

1. Olga Nevzorova, co-chair, Cand.Sc., Tatarstan Academy of Sciences, Institute of Applied Semiotics, Kazan, Russia
2. Rinat Gilmullin, Cand.Sc., Tatarstan Academy of Sciences, Institute of Applied Semiotics, Kazan, Russia
3. Ajrat Gatiatullin, Cand.Sc., Tatarstan Academy of Sciences, Institute of Applied Semiotics, Kazan, Russia
4. Alfiya Galieva, Cand.Sc., Tatarstan Academy of Sciences, Institute of Applied Semiotics, Kazan, Russia

Editors

1. Natalia Loukachevitch, Dr.Sc., Lomonosov Moscow State University, Research Computing Center, Moscow, Russia
2. Alexander Elizarov, Dr.Sc., Kazan (Volga Region) Federal University, N.I. Lobachevsky Institute of Mathematics and Mechanics, Kazan, Russia

Nonverbal Elements in Everyday Russian Speech: An Attempt at Categorization

Natalia Bogdanova-Beglarian[0000-0002-7652-0358] and Ekaterina Baeva[0000-0002-6045-1044]

Saint Petersburg State University
7/9 Universitetskaya Emb., St. Petersburg, 199034 Russia
{n.bogdanova,e.baeva}@spbu.ru

Abstract. The article provides an attempt at systematization of the elements of oral discourse which are not related to the text content but are nonetheless very frequent in everyday speech and thus essential for its understanding and decoding.

Nonverbal elements can be tracked almost in any type of spoken speech or any given speaker. Therefore it is essential to have a comprehensive classification which will enable researchers to deal with spoken speech data with more precision. Such elements include some filled hesitation pauses such as [ə:], [ə:m], [i:], [n], etc., nonverbal vocalizations like clicking, lip-smacking and squelching, as well as a number of other paralinguistic elements (voice qualifications such as laughing, sighing, coughing and so on).

The aim or the paper is to list various nonverbal elements in The Speech Corpus of the Russian Language (amounting to 1280 hours of recorded everyday Russian speech of more than 250 respondents and about 1000 of their interlocutors) and categorize them with regard to their pragmatic meaning. Nonverbal vocalizations usually tend to fill the hesitation pauses marking the so-called points of failure. Moreover, they often help to structure a text being produced and sometimes perform several functions simultaneously. While being hesitative, can also perform search functions (when a speaker searches his mind for a word, an expression or an idea to continue or complete an utterance), be a reflexive marker or as a discursive marker of the speech start or finale.

Keywords: modern Russian, everyday speech, nonverbal vocalizations, paralinguistic elements, speech corpus, hesitation phenomena

1 Introduction

It has been widely acknowledged that in contrast with written discourse, spoken speech has its own rules and therefore requires special research methods and approaches. In order to help describe and analyze contemporary Russian speech, three key elements have been drawn up [1].

Verbal elements are in the core of the semantic dimension of a text; they carry the principal meaning of a message. *Roughly verbal elements* are characterized by high frequency and high repetition; they help structure the text without actually being connected to its meaning. They are auxiliary parts of speech and parenthetical words. Moreover, to this category belong pragmatic markers, for example, verbal hesitatives of search ("*kak ego"* 'whatshisname'). The research of Russian pragmatic elements, if not sufficient, is definitely striving at the moment; we can consider the studies of K. L. Kiseleva's and D. Paillard's works [2, 3, 4] the pioneers of in-depth research of discourse words in Russian. Among others, there are works by G. Bolden [5, 6], T. Sherstinova [7], D. Dobrovolskij and L. Poppel [8, 9] dedicated to discursive pragmatic units in contemporary Russian speech. The studies focus primarily on "auxiliary" speech items. These pragmatic markers, as a rule, are characterized by significant weakening of their lexical and/or grammatical meaning. Nevertheless, they have an extremely high frequency, exceeding that of almost all content, textual units in spoken discourse.

Nonverbal elements of speech stand out in every utterance because they are rather frequent, yet they do not seem to bear any significance with regard to an utterance meaning. Apparently, being highly repetitive, they can structure and even pace a text without actually being of textual nature. These elements include hesitation pauses as a major part of spoken discourse.

While describing nonverbal communication in English, which usually implies visual information like gestures from face, eyes, hands and other body parts, D. Crystal [10] suggests dividing paralinguistic features into voice into *voice qualifiers* (such as whispery, breathy or creaky voice) and *voice qualifications* (like laugh, giggle, sob or cry). The latter group, together with physiological reflexes, belongs to non-word vocalizations that are termed *nonverbal vocalizations* [11, 12].

These elements have also been found rather frequent in everyday speech; however, their research in Russian speech has been devastatingly scarce. These non-verbal elements of the utterance are considered to be a type of speech malfunctions disrupting the smooth deployment of the speech (disfluencies) [13] and, as will be later shown the analysis of the corpus material, can be attributed

to non-verbal pragmatic markers because of the functions they perform in oral speech.

2 Nonverbal Elements in Speech

2.1 Hesitation Pauses

Non-verbal elements of speech, first and foremost, are hesitation pauses filled with non-phonemic sounds, or vocalizations. Pauses are considered to be an essential criterion for fluency rating and speech rate measurement. As a rule, pauses in speech are categorized into filled and unfilled, the former being hesitation particles like [ə:] or [ə:m] and the latter a simple silence. Filled pauses are an important indicator of speech fluency and therefore are widely investigated in studies dedicated to second language acquisition and mastering [14, 15, 16, 17].

It is the assumption that, in comparison to native speech, in non-native language the number of hesitations increases, which enforces the effect of slowing down and reduced fluency. However, it has been observed that "filled pauses" rarely occur in read speech [18].

2.2 Clicks

Clicks are usually described as phoneme realizations in some African languages [19] or as paralinguistic vocalizations, e.g. to signal disapproval or as sound imitation. Wright [20: 208] in her background research review offers a comprehensive summary of valences signaled by clicking in English: disapproval, annoyance, irritation, exasperation, impatience, regret, sympathy, and encouragement. She also emphasized that clicks usually occur in the vicinity of filled hesitation pauses which, in turn, would suggest formulation difficulties with regard to lexical or syntactic search, or signal new information [21].

Another recent discovery suggests that clicks are, presumably unintentionally, used as discourse markers indexing a new sequence in a conversation or before a word search. For example, J. Trouvain and Z. Malizs [22] investigated more than 300 apical clicks of an experienced speaker during a keynote address at an Interspeech conference. In turned out that the produced clicks occurred only in inter-speech intervals and were often combined with either hesitation particles like «uhm» or audible inhalation. Consequently, it is claimed that clicks are used as hesitation markers.

In Russian research clicks have rarely been identified and studied; however, some [23] list clicks among "artifacts", or short nonverbal elements which would be otherwise described among voice qualifications.

2.3 Voice qualifications

Physiological reflexes such as chewing noises, hiccup, coughing, yawning etc. are not usually considered communicative because they are not always under control of the speaker. However, some deliberate *vegetative sounds* (such as clearing the throat as indicating one's presence) can have pragmatic meaning and thus deserve further investigation [12].

Affect bursts [24] are vocalizations such as laughing, crying, screaming and many other short emotional non-speech expressions. More often than not, they are used deliberately and consciously. It is observed that affect bursts, even presented without context, can convey a clearly identifiable emotional meaning [25].

It is generally believed that nonverbal vocalizations occur more often in conversational speech than in monologues, reading at loud or other forms of controlled speaking. An analysis of six corpora of conversational speech [11] concluded that most common vocalizations were laughing and various types of breathing noises.

In addition to nonverbal vocalizations which can be investigated in several languages, there are those less widely acknowledged, e.g. lip-smack which is consistent with the Chinese language. It is a sound generated by pressing lips together and then opening them quickly, and it is considered to be a typical background event in Chinese spontaneous speech [26, 27]. However, Russian speakers have also been observed lip-smacking, albeit not very frequently, if compared to clearing your throat and coughing [23].

To summarize, we can see that nonverbal elements are very common in spontaneous speech. When conducting a thorough multi-level analysis of *verbal* spoken speech, one must detect and categorize its inherent *nonverbal* elements to help investigate and process more significant textual parts of any utterance. The current study is a part of ongoing research into pragmatics of spoken Russian, and based on this we now formulate the following research questions:

1) Which non-verbal elements can be found in everyday Russian speech?
2) How can we categorize them?

3 Research Method and Data

This study is conducted on the two modules of the Corpus of the Russian language: the corpus of Russian everyday speech "One Day of Speech" (the ORD corpus, containing mostly dialogic speech) [28] and "Balanced Annotated Text Collection" (SAT, containing monologic speech) [29].

The ORD corpus captures natural speech of native speakers (residents of St. Petersburg who speak Russian as their native language) and contains mostly everyday dialogues and polylogues, recorded using the method of continuous daily speech monitoring and recording. Each respondent provided about 8-14 hours of speech recordings which were then converted to the format of the corpus: PCM, 22050Hz, 16 bit, mono, while the original recordings had been stored in the archive. Next, the recordings were segmented into the so-called macroepisodes, in other words, fragments homogeneous in their communication settings which may include the place of communication, its settings, social roles of speakers or the activity they engage in. This segmentation was performed manually by qualified linguists who listen to the recordings and mark the boundaries between episodes.

The phonetic quality of each macroepisode is evaluated and measured in a 4-grade scale: 1 – the best quality, suitable for precise phonetic/prosody analysis, 2 – rather good quality, which is partially suitable for phonetic analysis, 3 – noisy recordings of intermediate and low quality, which are not suitable for phonetic analysis but are suitable enough for other aspects of research, and 4 – unintelligible conversations or remarks in extreme noise, which could not be understood without noise reduction techniques [30].

All data has been manually transcribed and later verified in ELAN [31], for the detailed principles of annotation and transcription see [28]. For data processing we used software specially designed for ORD, *Corrector* software utility (to correct possible technical errors in typescripts and to reveal potential mismatch in speaker/speech level in cases of overlapping utterances) and *Eafer* program (dissecting one-level transcript into a multi-level one). However, only selected macroepisodes of good quality or original content have so far been automatically processed and annotated on many levels. The work of comprehensive multilevel annotation of the whole corpus is obviously of large-scale nature and is still in progress.

Currently the corpus comprises 1250 hours of sound recordings, collected from 128 respondents and more than 1000 of their interlocutors, representing different social groups of St Petersburg, Russia, 2800 macroepisodes of communications, and more than 1 mln word usages in transcripts.

SAT, on the other hand, contains a less natural experimental speech. These are the monologues recorded from native speakers of different professional groups: doctors, lawyers, computer scientists, teachers of language and philosophy, various groups of students, incl. those majoring in language, and so on. SAT recordings are categorized into a series of typical communicative scenarios of everyday communication: reading, retelling, description of the image, story-telling. In addition to the speech of native Russian speakers, SAT also includes several blocks of L2 Russian speech by non-native speakers: American, French, Chinese, and Dutch. At the moment, the collection includes data obtained from 153 speakers and comprises 772 monologue texts, with total duration of 30 hours.

In brief, all data in the corpora is presented in both audio files and transcripts. An annotated ELAN file is presented in Fig. 1.

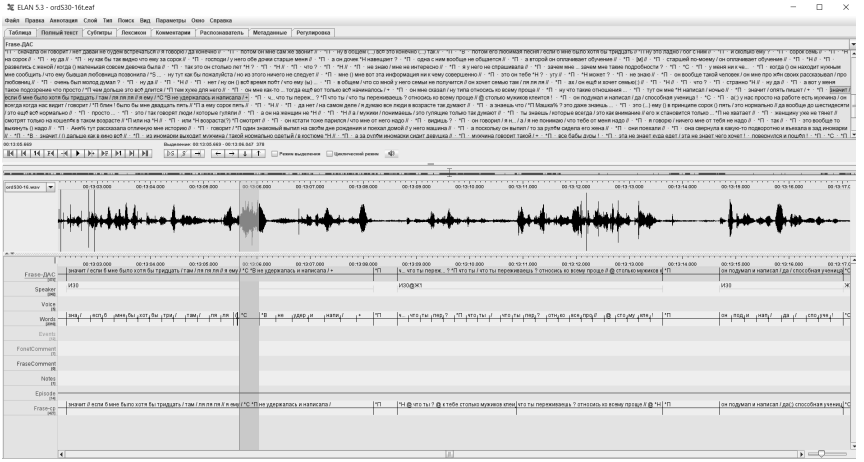


Fig. 1. An example on a multi-level annotated speech fragment in ELAN

It can be seen that there are certain symbols used in transcripts to mark non-verbal phenomena (**C, *B* and others). Most common symbols include *‘‘*П”* for a hesitation pause, “/” for a short utterance pause and “//” for a long pause marking the end of an utterance. Other symbols are introduced in their respective sections. All words and utterances are given in orthographic writing.

For the current classification study we explored both types of records and identified the phenomena we thought to be of non-verbal nature. Then we ana-

lyzed the phenomena and classified them into categories. It should be mentioned that so far the analysis is of qualitative nature rather than quantitative, principally because we aimed to create a classification to be proved or disproved in further research into spontaneous Russian speech.

4 Results and Discussion

4.1 Provisional Version of Nonverbal Vocalizations Inventory

Being the pioneers of comprehensive descriptions of Russian nonverbal vocalizations, we are faced with a series of debatable issues.

Firstly, we would aspire to compare and contrast our classification to those already existing in describing other languages, mainly English. So, one is expected to come up with an inventory similar or of the same nature, operating more or less similar terms and definitions. However – and here comes our second stumbling stone – there are abovementioned Russian studies of some, if not all of them, nonverbal elements in spoken Russian, and as native researchers we would not want to digress too far from our venerable colleagues.

As a result of our investigation, we have come up with a working theory for the typology of nonverbal elements in spoken Russian speech. In the corpus recordings managed to track the following elements:

- Hesitation pauses (filled and unfilled);
- Clicks;
- Lip-smacks;
- Noisy air intakes;
- Voice qualifications, or affect bursts.

This inventory serves as an exploratory one which is liable to undergo some alternations or refinements in the process of its validation on perhaps expanded speech material.

Hesitation pauses. Both types of hesitation pauses, filled and unfilled, can be found in the corpora, and they are rather frequent. Given that some subcorpora have been described in previous research, we can preview some quantitative data. For instance, in the SAT reading recordings (subgroup STU) there are 323 hesitation pauses [32].

There are different non-phonemic sounds that can fill a pause, predominantly [əː] or [əːm], [aː], [aːm], [mː]. In the Russian L2 speech of native Chinese speakers it was possible to trace sounds such as [y], [yn], [nː]. An example of a hesitation pause is provided below (see Fig. 2)

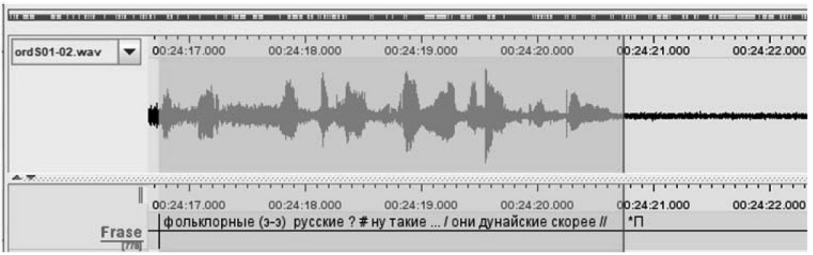


Fig. 2. An example of hesitation pauses in corpus data

The main function of filled pauses is hesitational search, either for a specific lexical unit or signal general speech formulation difficulty. More often than not, this search function seems to be accompanied by others. Let us consider some examples:

(1) i neskolko [m:] dvorovykh malchishek s treshchotkami (SAT, reading);

(2) Grigorij_Ivanovich [ə:] *Muromskij* [ə:].

Thus, in these examples, the speaker seems to hesitate before an archaic word uncommon for contemporary speech *"dvorovyj"* 'house serf' (1) and surname *"Muromskij"*. A previous study of this corpus data on lexical and syntactic level have previously suggested that there are markers of speech non-triviality which signal introducing some extraordinary, non-so-common verbal units, and they are often accompanied by hesitation pauses [33, 34].

The vocalizations often come together with other pragmatic markers, such as reflexive markers, markers of hesitation, or discursive markers. Thus, we may assume apparent polyfunctionality of vocalizations in oral discourse, with the hesitative-search character of almost all such elements as a given.

Clicks. As we already mentioned, clicks are not often specified in Russian research into spontaneous speech. In the current study, it was possible to locate clicking, marked as **Ц* in the corpus transcripts (see Fig. 3), in spoken Russian material, both in native and non-native speech.

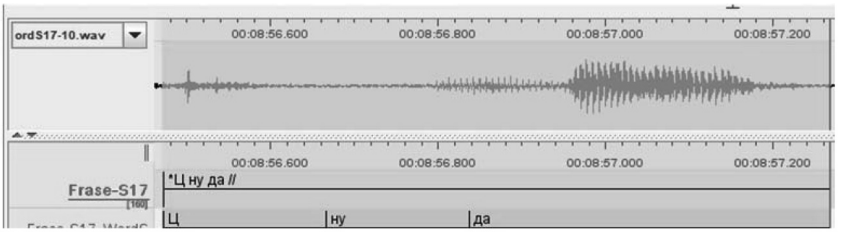


Fig. 3. An example of clicks in corpus data

However, more often than not clicks would be attributed to Chinese speakers. In most cases, clicks would definitely be of hesitative nature, and their primary function is word search, be it successful or not quite:

(3) *"nuzhno *Ц / za... / zanimatsya; ne ochen' [ə:] nravits'a(:)[əm:] *П *Ц khodit' v magaziny"* ("I have to study, I don't really like going shopping").

Similar to hesitation pauses, in clicks search function is also combined with the discursive start function: the speaker is found clicking at the start another fragment of his monologue. Again, there is some polyfunctionality of the non-verbal elements in oral communication.

Lip-smacks. Lip-smacks are marked as "*mp*" in our speech corpora findings and typescripts, mainly because of onomatopoeic reasons. These elements seem not very common, yet far from non-existent to be disqualified. It seems that the lip-smacks in spontaneous Russian, as well as all other types of hesitation phenomena, gives the speaker a short break for decision to continue speech or choosing the right word or expression and thus has a search function. Other markers of hesitation have been spotted in the vicinity: nonverbal sounds, prolongation of sounds, word breaks, parasite words and physical pauses, which further enhance their hesitational character.

Noisy air intakes. During this nonverbal vocalization a speaker draws in the air not through their nose, as it usually happens (including a situation when a deep breath is a hesitation pause by itself), but through the mouth, with the tip of the tongue at the front teeth, and between the lateral parts of the tongue and lateral teeth there is a gap through which the air passes. To an untrained ear it sounds like a noisy air intake. In some studies, it has been called *squelching* [34], and in the transcripts is marked as "sl":

Voice qualifications. There are several types of affect burns recorded and marked in the corpora typescripts, e.g. laughter (see Fig. 4), coughing, yawning, tutting, sneezing, etc. In Russian studies these are often called paralinguistic phenomena. Nevertheless, it seems that the nature of clicks, lip-smacks and noisy air intake would also attribute them as paralinguistic elements, which, however, do not carry much emotional significance.

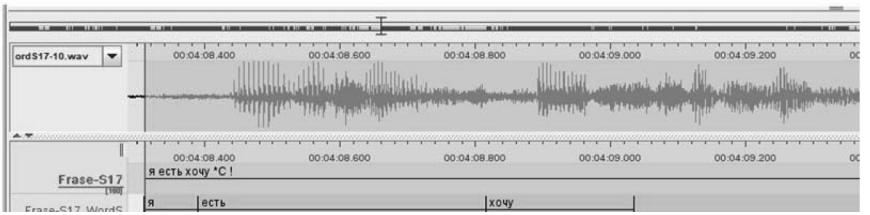


Fig. 4. An example of laughter in corpus data

4.2 Pragmatic Functions of Nonverbal Elements

Nonverbal elements of oral speech – in general, as a class of elements, and each separately – deserve a special functional description. But all our examples demonstrate their obvious hesitative nature, and also some polyfunctionality. For instance, clicks, like filled hesitation pauses, in addition to the search function may have the discursive start function. This fact may urge us to review the classification of nonverbal elements in the domain of pragmatics and thus consolidate some elements with regard mainly to their pragmatic function and not their phonetic execution.

At this point we may speak of three principal pragmatic functions: hesitation, search and reflection (often resulting in hypercorrection).

5 Conclusion

Our research shows that there are nonverbal elements in various types of oral discourse, in both monologues and polylogues. On the one hand, these elements do not claim to be significant, or verbal, and surely cannot be described as verbal. On the other hand, they have a definite pragmatic meaning and often help the speaker structure the speech he/she produces.

There are various approaches to categorization of nonverbal elements in spoken Russian speech, however, one cannot deny that these elements must be included in contemporary speech research, given their prolificacy.

The main function of nonverbal vocalizations we have found to be hesitative search, which is often intensified or modified by others: the functions of a discursive marker (start or final), a reflexive or a «non-trivial» marker. Corpus approach to the analysis of oral speech allows not only to identify all such «nonverbal» elements, but also to systematize them.

The findings may be used for many applied purposes: from teaching Russian in a foreign language audience to automatic speech recognition and linguistic expertise. Our study has been closely linked with fellow researchers' work into prosody and pragmatics, all of us striving to combine prosodic information with pragmatic annotation of communicative acts presented in the corpora. Further acoustical analysis of our identified categories of all non-verbal material, which is extremely common in spontaneous Russian speech, will allow for more precise automatic speech processing. This research, given its pragmatic aspect, is especially significant with regard to filled pauses recognition, as it has been observed that ASR systems tend to confuse filled pauses and backchannels, a functional distinction that humans need to be very good at pragmatically [35].

Acknowledgements

The presented research was supported by the Russian Science Foundation, project #18-18-00242 “Pragmatic Markers in Russian Everyday Speech”.

References

1. Russkij semanticheskij slovar’. Tolkovyj slovar’, sistematizirovannyj po klassam slov i znachenij. Tom 1: Slova ukazujushchie (mestoimenia). Slova imenujushchie: imena sushchestvitel’nye (Vs’o zhyvoe. Zeml’a. Kosmos) [Russian Semantic Dictionary. Dictionary, Systematized According to the Classes of Words and Meanings. Vol. 1: Pointing Words (Pronouns). Naming Words: Nouns (All Alive. Land. Cosmos)] / Shvedova, N,Ju. (ed.). Moscow: Azbukovnik Publ. 807 p. (1998).
2. Kiseleva, K. L., Paillard, D.: Diskursivnye slova russkogo yazyka: opyt kontekstno-semanticheskogo opisaniya [Discursive words of Russian: experience of contextual and semantic description]. Moskva: Metatekst Publ. (1998).
3. Kiseleva, K. L., Paillard, D. (eds.) Diskursivnye slova russkogo jazyka. Kontekstnoe var'irovanie i semantičeskoe edinstvo [Discourse words of Russian: contextual variation and semantic units]. Moskva: Azbukovnik (2003).
4. Paillard, D.: Discourse words in Russian. The case of voobšče and v obščem. Sprache und Datenverarbeitung, 30(1), 69–81. (2006).
5. Bolden, G. B.: Little words that matter: Discourse markers “so” and “oh” and the doing of other-attentiveness in social interaction. Journal of Communication, 56(4), 661–688 (2006).
6. Bolden, G. B.: Reopening Russian conversations: The discourse particle–to and the negotiation of interpersonal accountability in closings. Human Communication Research, 34(1), 99–136 (2008).
7. Sherstinova, T.: Macro episodes of Russian everyday oral communication: towards pragmatic annotation of the ORD speech corpus. In: International Conference on Speech and Computer 2015, 268–276. Springer, Cham. (2015).
8. Dobrovol’skij, D., & Pöppel, L.: Corpus perspectives on Russian discursive units: semantics, pragmatics, and contrastive analysis. In Yearbook of Corpus Linguistics and Pragmatics 2015, 223–241. Springer, Cham. (2015).
9. Dobrovolskij, D., & Pöppel, L.: The discursive construction дело в том, что and its parallels in other languages: A contrastive corpus study. In: The International Conference Dialogue 2016, Moscow, Russia, June 1–4, 2016, pp. 134–145. RSUH. (2016).
10. Crystal, D.: Prosodic Systems and Intonation in English. Cambridge: Cambridge University Press. (1969).
11. Trouvain, J., Truong, K.: Comparing non-verbal vocalizations in conversational

speech corpora, Proc. 4th Int'l Workshop on Corpora for Research on Emotion Sentiment & Social Signals, Istanbul, pp. 36–39 (2012).
12. Trouvain, J.: Laughing, breathing clicking – The prosody of nonverbal vocalisations. In: Proc. Speech Prosody, pp. 598–602 (2014).
13. Verdonik, D., Rojc, M., Stabej, M.: Annotating Discourse Markers in Spontaneous Speech Corpora on an Example for the Slovenian Language. In Language Resources and Evaluation, the Netherlands. Iss. 41 (2), 147–180 (2007).
14. Lennon, P.: Investigating fluency in EFL: A quantitative approach. Language learning, 40(3), 387–417 (1990).
15. De Jong, N. H., & Bosker, H. R.: Choosing a threshold for silent pauses to measure second language fluency. In: The 6th Workshop on Disfluency in Spontaneous Speech (DiSS), 17–20 (2013).
16. Rossiter, M. J.: Perceptions of L2 fluency by native and non-native speakers of English. Canadian Modern Language Review, 65(3), 395–412 (2009).
17. Götz, S.: Fluency in native and nonnative English speech (Studies on Corpus Linguistics, Vol. 53). John Benjamins Publishing (2013).
18. Cucchiarini, C., Strik, H., Boves, L:. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. Journal of the Acoustical Society of America 111/6, 2862–2873 (2002).
19. Maddieson, I.: Patterns of sounds. Cambridge: CUP (1984).
20. Wright, M.: On clicks in English talk-in-interaction. Journal of the International Phonetic Association 41(2), 207–229 (2011).
21. Wright, M.: Clicks as markers of new sequences in English conversation. 16th International Congress of the Phonetic Sciences (ICPhS XVI), Saarbrucken, 1069–1072 (2007).
22. Trouvain, J., & Malisz, Z.: Inter-speech clicks in an Interspeech keynote. In: INTERSPEECH 2016. International Speech Communication Association. Pp. 1397–1401 (2016).
23. Kip'atkova, I.S., Verkhodanova, O.V., Ronzhyn, A.L. Segmentacija paralingvisticheskikh fonacionnykh javlenij v spontannoj russkoj rechi [Segmentation of Paralinguistic Phonation Phenomena in Spontaneous Russian Speech] // Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia [Perm University Herald. Russian and Foreign Philology]. Iss. 2 (18), pp. 17–23 (2012). http://www.rfp.psu.ru/archive/2.2012/kipyatkova.pdf (accessed 10/10/2018).
24. Scherer, K. R.: Affect bursts. In: Emotions, 175–208. New York: Psychology Press. (2014).
25. Schröder, M.: Experimental study of affect bursts. Speech communication, 40(1-2), 99–116 (2003).
26. Li, A., Zheng, F., Byrne, W., Fung, P., Kamm, T., Liu, Y., ... & Chen, X.: CASS: A phonetically transcribed corpus of Mandarin spontaneous speech. In: Sixth International Conference on Spoken Language Processing (2000).

27. Li, Y., He, Q., Li, T., & Wang, W. A detection method of lip-smack in spontaneous speech. In: Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on. Pp. 292-297. IEEE. (2008, July).
28. Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNAI, vol. 5729. Springer, Berlin-Heidelberg, pp. 250–257 (2009).
29. Bogdanova-Beglarian, N.V., Sherstinova, T.Ju., Zajdes, K.D.: Korpus "Sbalansirovannaja annotirovannaja tekstoteka": metodika mnogourovnevogo analiza russkoj monologicheskoj rechi [The corpus "Balanced Annotated Text Collection": Method of Multilevel Analysis of Russian Monological Speech] // Analiz razgovornoj rechi (AR3-2017): trudy sed'mogo mezhdisciplinarnogo seminara [Analysis of Spoken Russian Speech (AR3-2017): Proceedings of the 7th Interdisciplinary Seminar] / Kocharov, D. A., Skrelin, P. A. (eds). St. Petersburg: Polytekhnica-print Publ., 8–13 (2017).
30. Sherstinova, T.: The structure of the ORD speech corpus of Russian everyday communication. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 258–265. Springer, Heidelberg (2009).
31. Hellwig, B., Van Uytvanck, D., Hulsbosch, M., et al.: ELAN – Linguistic Annotator. Version 5.0.0-alfa [in:]. http://www.mpi.nl/corpus/html/elan/. Accessed 28 Oct 2018.
32. Baeva, E.M.: Hezitacionnye javlenija v ustnyh monologah nizkoj stepeni spontannosti [Hesitation phenomena in spoken Russian speech of low spontaneity]. Kommunikativnye issledovania [Communicative Studies]. Iss. 1 (15). pp. 75–84 (2018). DOI: 10.25513/2413-6182.2018.1.75-84
33. Bogdanova-Beglarian, N.V.: «Netrivial'noe» v povsenevnoj ustnoj kommunikacii: opyt sistematizacii [«Non-trivial» in Everyday Communication: An Attempt of Systematization // Kommunikativnye issledovania [Communicative Studies]. Iss. 4 (14), pp. 9–30 (2017). http://www.com-studies.org/images/magazine/2017/4_14_2017.pdf (accessed: 10/10/2018)
34. Chen, Ch.: Hezitacii v russkoj ustnoj rechi nositelej kitajskogo jazyka [Hesitations in Russian Oral Speech of Native Chinese Speakers]. PhD Thesis. St. Petersburg. 205 p. (2018) https://disser.spbu.ru/files/disser2/disser/850r4wMfLM.pdf (accessed: 10/10/2018)
35. Stolcke, A., Droppo J.: Comparing human and machine errors in conversational speech transcription. In: Proc. Interspeech 2017, pp. 137–141 (2017).

Analysis of dynamics of the number of syntactic dependencies in Russian and English using Google Books Ngram

Vladimir V. Bochkarev[1] Valery D. Solovyev[2] and Anna V. Shevlyakova[3]

[1] Kazan Federal University, Kazan, Russia
vbochkarev@mail.ru
[2] Kazan Federal University, Kazan, Russia
maki.solovyev@mail.ru
[3] Kazan Federal University, Kazan, Russia
anna_ling@mail.ru

Abstract. The work examines the dynamics of the number of syntactic dependencies and 2-grams in Russian and English using the Google Books Ngram diachronic corpus. We counted the total number of 2-grams and syntactic dependencies detected in Google Books Books Ngram at least once in a given year, as well as stable dependencies, which value of pointwise mutual information is above a given threshold. The effective number of dependencies expressed through the perplexity of 2-gram frequency distributions was also calculated. This value is a characteristic number of frequently used word combinations. It was found that quantitatively unchanged core and rapidly growing periphery can be distinguished among the syntactic dependencies of words. It was possible to obtain an estimate of the growth rate of the effective number of syntactic dependencies in the Russian language. The estimate shows that doubling of the effective number of dependencies occurs approximately every 250 years if the corpus size stays unchanged.

Keywords: Google Books Ngram, syntactic dependencies, computational linguistics, correlation models, linguistic databases.

1 Introduction

Emergence of extra-large text corpora and development of new algorithms and methods of linguistic research opens up broad opportunities for studying dynamic processes occurring in a language, and allows us to trace evolution of language phenomena.

Computer processing of large arrays of language data makes it possible to quantify the dynamics of lexicon and development of intralingual relations, classification and clustering of vocabulary. One of the largest corpora of texts is the Google Books library [1, 2]. It includes more than 8 million of digitized books written in 8 languages and is currently the largest digital text resource. The oldest books included in the corpus were written in the 1500s, and the latest book was published in 2009. The Google Books Ngram services allow frequency analysis of word usage and visualization of the data.

Performing a quantitative analysis of text corpora, researchers solve various problems concerning language complexity [3], interrelations between language and culture (even the special term "culturomics" was introduced) [1], try to detect regularities of emergence and functioning of linguistic units and evolution of grammar. The article [3], in which the growing number of unique phrases in the English language was studied seems to be the most interesting in the context of our work. The author explains that increase in the number of word combinations is due to increasing complexity of culture. Meanwhile, the size of the Google Books Ngram corpus constantly increases (see Figure 1). The corpus growth, by itself, in accordance with Heaps' law, should lead to growth in the number of unique word combinations. The empirical Heaps' law describes the dependence of the number of unique words in a text on the size (length) of this text and states that the number of these words is connected by a power dependence with the size of the text [4, 5]. Despite the fact that the classical formulation of Heaps' law speaks only about the number of unique words, the same applies to the number of word combinations and syntactic dependencies [6]. Also, a certain disadvantage of Juola's work is that all of the conclusions are based on the analysis of the English corpus only.

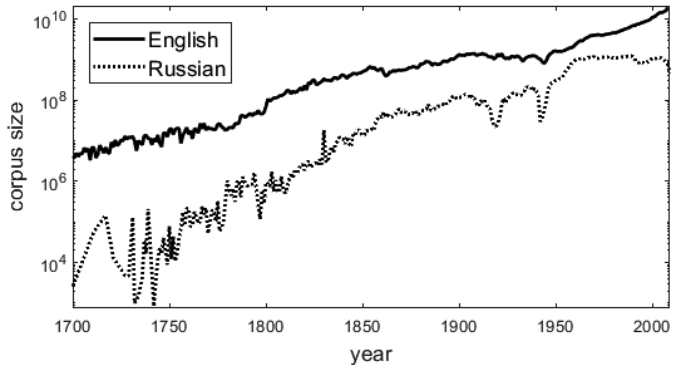


Fig. 1. Size of the common English and Russian sub-corpora included in Google Books Ngram (number of words)

Taking into account the conclusions [3], we set out a goal to analyse the dynamics of the number of syntactic dependencies and 2-grams. A priori, it can be expected that the number of such word relationships increases over time due to two factors: 1) increasing complexity of human culture [7, 8] and emergence of new words providing increase in the number of semantic connections and syntactic dependencies; 2) metaphorization processes, which also increase the number of relationships between words. Also, the number of 2-grams and syntactic dependencies detected in the corpus grows due to increase of the corpus size. The study objective was to identify how the number of 2-grams and syntactic dependencies increases with time, as well as to trace the impact of each of these factors. The Russian and English text corpora, which belong to the diachronic corpus Google Books Ngram, were studied.

2 Data and Methods

The common corpus of the English language and the corpus of the Russian language, which are a part of Google Books Ngram, were analysed.

Raw data are available for download on the project page (https://books.google.com/ngrams/). They contain information on frequency of use of words and n-grams (2-, 3-, 4- and 5-grams) in the books presented in the Google Books electronic library for each year. In our work, we used a base of frequencies of 2 grams, that is, pairs of words which, directly go one after another in the sentence.

A distinctive feature of the version of the 2012-year corpus is the presence of a base of frequencies of syntactic dependencies. Syntactic dependencies are understood as pairwise relationships between words in the same sentence. One of the words is a head, another one is a modifier. Such dependency relations are independent of word order, even though there are often intervening words between the head and the modifier. The data on frequencies of syntactic dependencies available in the Google Books Ngram corpus were also used in this work.

Thus, the term “2-gram” is used in our work when we describe pairs of words, which directly go one after another in the sentence. The term “syntactic dependencies” is used for head-modifier pairwise relationships between two words in a sentence. We study the number of different 2-grams and pairs of words being in a syntactic dependency.

Preliminary data processing was performed before the study. First, we did`t make a distinction between words that differ in case. Accordingly, 2-grams and syntactic dependencies, containing words that differ in case, were considered identical. Secondly, only vocabulary 1-grams were selected. 1-grams are

understood to be words composed only of letters of the corresponding alphabet and, possibly, one apostrophe. If not taking into account differences in case, there were 5096 thousand (out of the total number of 8256 thousand) of such 1-grams found in the common English corpus. Accordingly, 4091 thousand 1-grams out of a total number of 5096 thousand 1-grams were selected for the Russian corpus. To normalize and calculate relative frequencies, the number of vocabulary 1-grams was calculated for each year (unlike Google Books Ngram Viewer, where normalization is performed for the total number of 1-grams). Parts of speech are marked in the 2012 version of the database. However, in many cases, parts of speech are determined improperly, which can cause incorrect conclusions based on such data. Therefore, the method introduced in [9] was used. It says that if the number of word forms corresponding to a certain part of speech does not exceed 1% of the total frequency of use of this word form, such word forms should be rejected and not used in further analysis. During the second stage of the survey, 2-grams consisted of the selected 1-grams were analysed.

The analysis was based on the following principles. Many researches attempt to determine the number of word combinations in the language. The easiest way to do it is to count the number of different word combinations in a corpus in a given year. To analyze the number of pairs of words forming dependencies, we counted the total number of 2-grams and syntactic dependencies marked in the Google Books Ngram database at least once in a given year. However, this method has some drawbacks. The first drawback is that a large amount of word pairs located next to each other in a sentence but not forming a dependency is counted. The second drawback is that, according to the authors of the Google Books Ngram project, approximately 30% of unique word forms contained in the database result from misprints. These factors cause an even more significant overestimation of the number of 2-grams and syntactic dependencies. The third drawback is that empirical frequencies of rare words, which are in the majority in the base, highly fluctuate, which also leads to large errors in estimation of the number of 2-grams and syntactic dependencies. Two approaches were used to reduce the impact of these factors. The first one is the following. Not all 2-grams and syntactic dependencies were counted but only frequently used ones, which are in a certain associative connection and are called collocations. Usually collocations are understood as word combinations, where words a located next to each other. However, some researches consider that stable syntactic dependencies can also be called collocations [10].

A value called pointwise mutual information in computational linguistics [11, 12] was used as a measure of associative connection. This value is expressed by the formula:

$$\mathrm{MI} = \log_2 \frac{f_{12}}{f_1 f_2} \tag{1}$$

Here f_{12} is a relative frequency of the word combination, and f_1 and f_2 are relative frequencies of the words, which form the word combination. As can be seen from the formula, the MI value shows to what extent the word combination is found more often in a text or a corpus than in a random text of the same size with an independent choice of words. The selection was carried out according to the value of the MI, which is 0, 3, 6, 9 for a number of threshold values of this quantity. The calculation results for the English and Russian languages are shown in Figure 2.

The second possible solution may be to count the number of word combinations with regard to their informational content. We can used such characteristic of frequency distribution as perplexity [13]. The effective number of syntactic dependencies (2-grams) numerically equal to the perplexity of their frequency distribution was introduced:

$$N_{eff} = 2^h \tag{2}$$

Here h is the entropy of the frequency distribution, calculated by the formula:

$$h = -\sum_i f_i \log_2 f_i \tag{3}$$

where f_i is the frequency of the i-th 2-gram (or syntactic dependency). The introduced value shows the number of frequently used syntactic dependencies

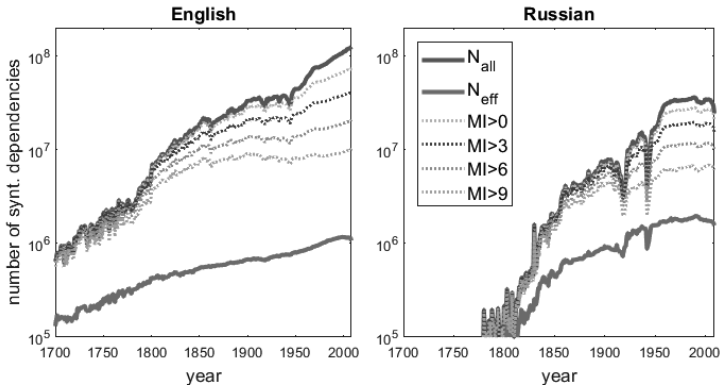


Fig. 2. The number of syntactic dependencies in Russian and English in 1700-2008. The total number of syntactic dependencies, the effective number of syntactic dependencies (perplexity) and the number of syntactic dependencies with MI above the given threshold are shown

(2-grams), taking into account their role in the information exchange. Our approach is close to that used in [3]. However, using perplexity instead of entropy allows us to present the results more vividly, as well as to make comparisons with estimates obtained by other methods. The dynamics of the effective number of syntactic dependencies of both languages is also shown in Figure 2.

3 Results

As can be seen from Figure 2, the total number of syntactic dependencies in both languages is growing rapidly. At that, the growth rate in different periods changes significantly, the curve responds to various historical events, primarily to wars and revolutions. If we restrict ourselves to stable syntactic dependencies, the curve qualitatively retains its character. However, it shows a slightly lower growth rate. The number of syntactic dependencies with high MI values grows very slowly. All this is true for the number of 2-grams.

Comparing figures 1 and 2, it can be seen that the curves of the total number of syntactic dependencies are similar to the graphs of the corpora size. This observation can be quantified. Table 1 shows the values of the Spearman correlation coefficients between the corpus size and the number of syntactic dependencies (the total number of syntactic dependencies and the number of only stable syntactic dependencies) in English and Russian.

The correlation coefficients will not change in any noticeable way if they are calculated using the limited intervals of 1700-2008 or 1750-2008. Thus, the compared values show a high level of statistical connection, especially for the Russian language.

Table 1. Spearman's correlation coefficient between the corpus size and the number of syntactic dependencies in Russian and English

	English	Russian
Total number of syntactic dependencies	0.890	0.974
Number of syntactic dependencies with MI>0	0.860	0.972

The graph of the effective number of syntactic dependencies has a different character. The curve is much more regular and smooth and responds insignificantly to historical events. The size of the English corpus is substantially larger than the Russian one. It contains approximately 470 billion of words and the Russian corpus includes only 67 billion of words. The English corpus shows no reaction to

historical events, and the graph of the effective number of syntactic dependencies can be well described by an exponential dependence (in a logarithmic coordinate system – a linear dependence). The smooth exponential growth of the effective number of syntactic dependencies in the English language is accelerated only after about 1950, which may be a manifestation of globalization processes. It is indisputable that by the end of the 20th century, English becomes the leading world language. Its influence on the processes of international economic, political and cultural integration proceed is great. English has also become the second mother-tongue for many people and develops very fast. The total number of syntactic dependencies in the English language is higher than in Russian, which is a manifestation of Heaps' law.

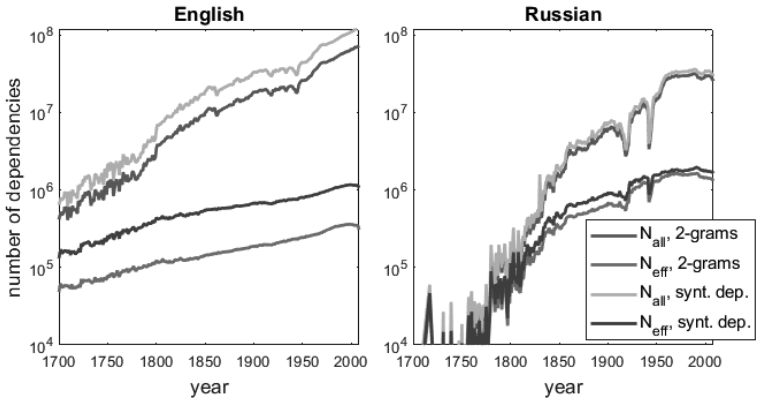


Fig. 3. The number of syntactic dependencies and 2-grams in Russian and English in 1700-2008

At that, the Russian language has more effective syntactic dependencies than English, which can probably be due to more complicated morphology and word-formation. Thus, applying such indicator as the effective number of syntactic dependencies allows us to perform less subjective comparative analysis of language processes using corpora of various sizes.

Figure 3 shows the number of syntactic dependencies and 2-grams in Russian and English. Both the total and effective number of syntactic dependencies and 2-grams are compared. Attention should be paid to the fact that the ratio of the number of syntactic dependencies to the number of 2-grams is significantly larger in the English language. Probably, this can be due to the fact that the word order in the Russian language is not fixed. As a result, a larger number of 2-grams can

be formed. This may also be due to some features of syntactic analysers used for creating a corpus. Otherwise, as can be seen from Figure 3, the number of syntactic dependencies and the number of 2-grams change over time in a similar way.

As it was stated above, increase in the number of syntactic dependencies and 2-grams can be due to growing complexity of culture, increase of a corpus size and metaphorization processes, which cause emergence of new words. Influence of each factor was investigated in the work.

To level the effect of a simple increase in the number of new words, one can count the number of word combinations and syntactic dependencies, which are comprised only of a fixed set of words belonging to the lexicon core. There are various approaches to the problem of determining the lexicon core [14]. To solve the problems mentioned in the article, it seems natural to use the method proposed in [15], according to which we select words recorded in the corpus each year from a certain period. There are approximately 37 thousand of words, which appeared in the common corpus of English each year between 1750 and 2008 (the amount of annual text data was insufficient before that time). Russian words appeared in the corpus every year between 1920 and 2008 were selected. To avoid difficulties associated with the impact of the 1918 spelling reform, the analysis was performed for the stated period. To make the conditions of comparison more equal for both languages, Russian words, which appeared each year at least 10 times, were selected. There were 80 thousand of words, which satisfied the required conditions.

Figure 4 shows the comparison between the change in the effective number (see formulae (2, 3)) of syntactic dependencies of all words and words, which belong to the lexicon core. The number of syntactic dependencies between words from the core grows much slower than that between all words. At that, the number of syntactic dependencies between core words has not grown since 1850. However, a small increase is observed only after 1960. Thus, the growth in the number of syntactic dependencies is largely due to the emergence of new syntactic dependencies for words from the lexicon periphery, as well as syntactic dependencies between words from the lexicon core and periphery.

Let us further consider how the total number of syntactic dependencies and 2-grams varies depending on the number of words in the lexicon. Assuming the validity of Heaps' law for both the number of words and the number of syntactic dependencies, it can be said that there should be power dependence between these quantities. Figure 5 shows the change in the number of syntactic dependencies and 2-grams depending on the number of unique words in English and Russian. Each point on the graph corresponds to the number of words and the number of

syntactic dependencies (2-grams) detected in the corpus in a given year (in the period 1505–2008 for the English language

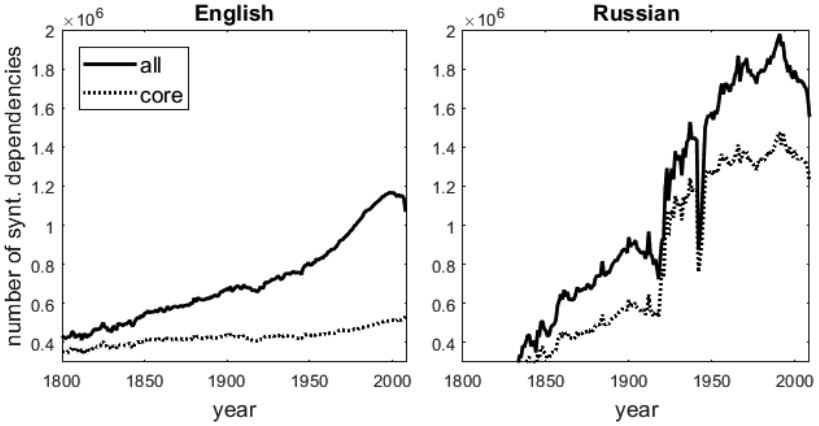


Fig. 4. Effective number of syntactic dependencies in English and Russian (both for the entire lexicon, and for the lexicon core)

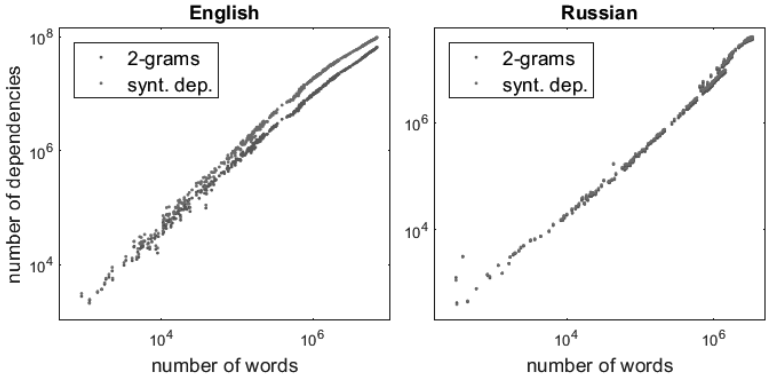


Fig. 5. Dependence of the number of syntactic dependencies and 2-grams detected in the corpus on the number of words in the lexicon

Dependences shown in Figure 5 are close to a power law, however, differences are also observed. It can be seen that the slope of the graph slightly differs in different areas. These differences may be due to variations of Heaps' exponent with time described in [14]. Performing approximation of the empirical data by a power law on the most important area (for the number of

words more than $1.5 \cdot 10^6$), we obtain the value of the power exponent for syntactic dependencies in the English language is equal to 1.174 (for word combinations - 1.169). That is, the number of syntactic dependencies per word grows slowly as the language becomes more complex. However, if we restrict ourselves only to stable syntactic dependencies with MI> 0, the power exponent for the number of syntactic dependencies will be 0.793 (0.815 for word combinations). Thus, the number of stable syntactic dependencies and word combinations per word falls. In both cases, the difference in the values of the power exponent for the number of syntactic dependencies and the number of word combinations is not significant. The difference of the power exponents from 1 is small, however, it can be important, since many growth models of complex networks predict proportionality of the number of network vertices (in our case, vertices are words) and the number of dependencies (in our case, syntactic dependencies) [16].

As for the Russian language, the power exponent for the number of syntactic dependencies is 1.097 (1.11 for the number of phrases) and equals 0.955 for the number of stable syntactic links with MI> 0 (0.96 for the number of stable phrases) under similar conditions. It should be noted that it is more difficult to find a linear segment for the Russian language in Figure 5. Therefore, these results are less reliable. Nevertheless, they are in good agreement with the estimates obtained for the English corpus.

Let us estimate quantitatively the degree of statistical connection between the number of unique words and the number of syntactic dependencies. Table 2 shows the Spearman correlation coefficients between these values for the English and Russian languages.

Table 2. Spearman's correlation coefficient between the number of unique words and the number of syntactic dependencies in Russian and English

	English	Russian
Total number of syntactic dependencies	0.999	0.981
Number of syntactic dependencies with MI>0	0.994	0.983

Comparing with the values given in table 1, it can be seen that the statistical connection between the number of syntactic dependencies and the number of words is even more significant than connection with the corpus size. A more significant increase is observed for the English language. This may be due to the fact that the saturation effect described in [14] (which is more pronounced for a larger English corpus) weakens the dependence of the number of syntactic dependencies and 2-grams on the corpus size.

If the corpus stays unchanged, the number of syntactic dependencies changes in the following way. The number of books represented in the Google Books Ngram Russian sub-corpus varies greatly in different years. The largest amount of books belongs to the period 1960-1991. From 65 to 80 thousand of volumes were published annually in the USSR in this period, and the corpus contains approximately 10 thousand volumes published each year (or 1-1.25 billion words), that is, at least 12% of all published books. Thus, there is a 31-year time period during which the size of the corpus varied within small limits. This provides an opportunity to assess the rate of growth of the number of syntactic dependencies directly, without taking into account the impact associated with the growth of the corpus size.

Figure 6 shows the change in the effective number of syntactic dependencies in the Russian language in the target period. The dotted line shows approximation of the series of the number of syntactic dependencies by exponential dependence using only the data from the period 1960-1990. The exponent rate was $2.74 \cdot 10^{-3}$, which corresponds to a doubling of the effective number of syntactic dependencies within 253 years.

Fig. 6. Change in the effective number of syntactic dependencies in the Russian language in 1955-1995

There is no period when the English language corpus size changes insignificantly. Nevertheless, if we approximate the curve of the effective number of syntactic dependencies in English in the same interval 1960-1991, the value of the exponent will be $9.36 \cdot 10^{-3}$, which corresponds to doubling of the number of syntactic

dependencies within 74 years. If we take the 1850-1950 data (see Figure 4), the exponent will be estimated as $3.49 \cdot 10^{-3}$, which corresponds to doubling of the number of syntactic dependencies within 199 years. The latter value is close enough to the above estimate obtained for the Russian language.

4 Conclusion

The number of 2-grams and syntactic dependencies detected in the Google Books Ngram corpus grows extremely rapidly. It increased by a factor of 160 for the common corpus of English and by a factor of 66 for the Russian corpus over the period 1800-2000. It is obvious that most of this growth is associated not with increase of language complexity, but with an extensive increase of the corpus size. To study the factors causing language complexity, it is more convenient to use not the total number of syntactic dependencies and 2-grams, but the number of stable syntactic dependencies and 2-grams (with MI above a given threshold) or their effective number (calculated as perplexity of frequency distribution). The latter characteristic demonstrates much smoother and regular change compared to the total number of the studied word relationships. The curve of the effective number of syntactic dependencies and 2-grams practically does not respond to historical events and, when calculated using the entire English vocabulary, it shows growth, according to the law close to exponential. However, the effective number of syntactic dependencies and 2-grams detected in the corpus each year over a fairly long time interval (1750–2008 for English and 1920–2008 for Russian) changes very slowly. This can indicate that quantitatively unchanged core and rapidly growing periphery can be distinguished among the syntactic dependencies of words.

It was found that the effects associated with the emergence of new words dominate among the factors influencing the growth in the number of syntactic dependencies and 2-grams. The dependence of the total number of syntactic dependencies and 2-grams on the number of unique words is close to a power law. It is clear that the power law should be considered only as some approximation of the empirical data. However, it should be noted that the power dependence in this case corresponds better to the empirical data than to the dependence of the number of syntactic dependencies and 2-grams on the corpus size (which is expected in accordance with Heaps' law). The same is true for the number of stable dependencies (with MI> 0). At that, the power exponents are slightly greater than 1 (1.1-1.17) for the total number of syntactic dependencies and 2-grams and less than 1 (0.79-0.96) for the number of only stable syntactic

dependencies and 2-grams for the studied languages. These facts should be taken into account when building models of growth of a network of syntactic dependencies in natural languages.

It was possible to obtain an estimate of the growth rate of the effective number of syntactic dependencies in the Russian language. If the corpus size stays unchanged, doubling of the effective number of syntactic dependencies should occur in 250 years. The effective number of syntactic dependencies in the English language is characterized by similar growth rates over a long period of time. However, their number increases approximately after 1950. This can be due to the fact that English is a global language.

Acknowledgements.

This research was financially supported by the Russian Government Program of Competitive Growth of Kazan Federal University, state assignment of Ministry of Education and Science, grant agreement № 34.5517.2017/6.7 and by RFBR, grant № 17-29-09163.

References

1. Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., et al.: Quantitative analysis of culture using millions of digitized books. Science 331(6014), 176-182 (2011).
2. Lin, Y., Michel, J.-B., Aiden, E.L., Orwant, J., Brockman, W., Petrov, S.: Syntactic Annotations for the Google Books Ngram Corpus. In: Li, H., Lin, C.-Y., Osborne, M., Lee, G.G., Park, J.C. (eds.) 50th Annual Meeting of the Association for Computational Linguistics 2012, Proceedings of the Conference, vol. 2, 238-242. Association for Computational Linguistics, Jeju Island, Korea (2012).
3. Juola, P.: Using the Google N-Gram corpus to measure cultural complexity. Literary and Linguistic Computing 28(4), 668–675 (2013).
4. Gerlach, M., Altmann, E.G.: Stochastic Model for the Vocabulary Growth in Natural Languages. Physical Review X 10(3), 021006 (2013).
5. Bochkarev, V.V., Lerner, E.Yu., Shevlyakova, A.V.: Deviations in the Zipf and Heaps laws in natural languages. Journal of Physics: Conference Series 490(1), 012009 (2014).
6. Williams, J. R., Lessard, P.R., Desu, S., Clark, E.M., Bagrow, J.P., Danforth, C.M., Dodds, P.: Zipf's law holds for phrases, not words. Scientific Reports 5, 12209 (2015).
7. Chick, G.: Cultural complexity: The concept and its measurement. Cross-Cultural Research 31, 275–307 (1997).

8. Arbib, M.A.: How the Brain Got Language: The Mirror System Hypothesis. Oxford University Press, Oxford (2012).
9. Bochkarev, V.V., Solovyev, V.D., Wichmann, S.: Universals versus historical contingencies in lexical evolution. J. R. Soc. Interface 11, 20140841 (2014).
10. Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., et. al.: Syntactic dependency-based N-grams as classification features. LNAI 7630, 1-11 (2012).
11. Fano, R.M.: Transmission of Information: A Statistical Theory of Communications. M.I.T. Press, Cambridge Mass. (1961).
12. Church, K., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16(1), 22–29 (1990).
13. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Lai, J.C., Mercer, R.L.: An Estimate of an Upper Bound for the Entropy of English. Journal of Computational Linguistics 18(1), 31-40 (1992).
14. Solovyev, V.D., Bochkarev, V.V., Shevlyakova, A.V.: Dynamics of core of language vocabulary. CEUR Workshop Proceedings 1886, 122-129 (2016).
15. Buntinx, V., Bornet, C., Kaplan, F.: Studying Linguistic Changes over 200 Years of Newspapers through Resilient Words Analysis. Frontiers in Digital Humanities 4, 1-10 (2017).
16. Durrett, R.: Random Graph Dynamics. Cambridge University Press, Cambridge (2007).

A Neural Network Approach to Morphological Disambiguation Based on the LSTM Architecture in the National Corpus of the Tatar Language

Rinat Gilmullin[1], Bulat Khakimov[1,2], Ramil Gataullin[1]

[1] Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan, Russia
[2] Kazan Federal University, Kazan, Russia
rinatgilmullin@gmail.com

Abstract. This paper presents the results of experiments on morphological disambiguation in the National corpus of the Tatar language "Tugan tel". The experiments were conducted using the LSTM based neural network model. The tagged socio-political sub-corpus of the National corpus of the Tatar language "Tugan tel" with a volume of 2,4 million words was used as training data. Experiments have shown that LSTM models are language-independent and can be applied to the Tatar language too. The results for Tatar are on a comparable level with those for other agglutinative languages, such as Hungarian and Turkish.

Keywords. Morphological disambiguation, Tatar language, Tatar National Corpus, corpus data, morphological tagging, LSTM, neural architectures

1 Introduction

Morphological disambiguation is one of the main tasks of automatic natural language processing. Its results can be used to improve accuracy and quality of the methods used in such tasks as text classification and clustering, machine translation, and information retrieval.

The complexity and peculiarities of morphological disambiguation vary for each particular language. For example, for English with its poor morphology and rigid word order in the sentence, the morphological disambiguation, as a rule, is reduced to the task of POS tagging and is based on rather simple methods. In Russian, morphological ambiguity is not so salient as in English, but, nevertheless, it is inherent. Free word order in Russian adds complexity to the task. In the Tatar language, as in other agglutinative languages of the Turkic

group, morphemes are the most important meaningful language units that carry both semantic and syntactic information. With a theoretically unlimited number of morphemes attached to the stem, morphological ambiguity takes on various forms, which greatly complicates the disambiguation.

Up to now, a basic paradigm of methods for disambiguation has been formed [1]. This includes the rule-based methods [2,3], machine learning methods based on the probabilistic models [4,5], and hybrid methods [6,7,8]. Developing the National corpus of the Tatar language "Tugan tel" (http://tugantel.tatar/) and the socio-political sub-corpus with manual morphological disambiguation made it possible to study this problem using statistical methods based on machine learning [3,8].

Analysis of open source codes developed for this task over the past few years has shown that one of the most effective tools is PurePos 2.0 [6] which implements a hybrid model based on hidden Markov models, as well as a neural network model based on recurrent neural networks with long short-term memory LSTM [5]. Hidden Markov model is a process model in which a process is considered a Markov process, and it is not known what state the system is in (its states are hidden), but each state can produce, with some probability, an event that can be observed. In other words, the Markov process with unknown parameters is studied, and the task is to recognize these unknown parameters basing on observables. The results of recognizing POS tags of Tatar words showed an accuracy of 97% [8].

Another approach that rather successfully solves the problem of morphological ambiguity is based on a recurrent neural network with a long short-term memory (LSTM) [9,10]. In [5], the results (see Table 1) of applying this approach to Turkish, Russian and Arabic are given.

Table 1. Results of experiments using the LSTM neural network architecture for morphological disambiguation

Language	**Turkish**		**Russian**		**Arabic**	
	% from ambiguous words	% from all tokens	% from ambiguous words	% from all tokens	% from ambiguous words	% from all tokens
Without context (baseline)	88.65	95.45	64.97	88.58	72.22	78.06
Local context	89.18	95.67	71.56	90.72	80.10	84.29
Whole sentence (surface form)	91.03	96.41	69.49	90.05	86.45	88.95
Left-to-Right	90.50	96.19	68.55	89.75	89.30	91.27
CRF	90.24	96.09	72.78	91.13	-	-

The analysis of the used context size in [5] deserves a special attention. The authors compared different sizes and types of contexts and experimentally revealed the most appropriate type for each language. It turned out that for the Turkish language it is sufficient to construct vectors based on surface word forms without explicitly defining their morphological features, but using all the words in the sentence. Whereas for Russian, agreement in gender, number and case is important, which in turn requires not only surface word forms, but also their morphological features in the context. At the same time, optimization based on the conditional random field method (CRF) helps to achieve better results (disambiguation accuracy 91.13%). The situation is similar with the Arabic language, when surface word forms are not enough for full disambiguation. This can be explained by the fact that in Arabic the level of ambiguity is higher than Turkish. If, for example, in Turkish, on average, there are 2.81 parsing options per word, and in Russian 5.81, then in Arabic there are 11.31. Therefore, for correct model training, a completely disambiguated tagged context is required.

This article describes the results of applying the neural network model based on the LSTM architecture to morphological disambiguation in the National corpus of the Tatar language.

2 The Tatar Language

The Tatar language belongs to the Turkic group that forms a subfamily of Altaic languages. It is spoken in West-central Russia (in the Volga region) and in the southern parts of Siberia. The number of Tatars in Russia in 2010 was 5,31 million people [9]. In 2013, the existing language classifications [12, 13] described Tatar as an under-resourced language.

3 LSTM model for morphological disambiguation

Model training requires tagged disambiguated texts. The method supposes that each parse of an ambiguous word and its context is juxtaposed with vectors. In the first case, the vector is based on its lemma and morphological features, and in the second case, on the surface forms of the surrounding words; in addition, the vector can be expanded by morphological features. Here, the context is not limited to several words of the immediate vicinity of words and can reach the size of the entire sentence. After that, on the basis of the resulting pair of vectors, the distribution of conditional probabilities is constructed; from these the most probable parse is selected as the correct one.

According to [5], the LSTM model is designed to build a vector representation of an ambiguous word (vectors are constructed on the basis of the lemma and morphological features of each of the alternatives, then they are united into R matrix and the surrounding context (indicated by h vector). After using the *softmax* function on the product of R matrix and h vector, the distribution of probabilities of each parsing option in this particular context is constructed, on the basis of which morphological ambiguity is resolved in favor of the most likely alternative:

$$p(y_t = a|x) = softmax(R_{x_t} vo\ h_t)$$

3.1 Vector representation of the ambiguous word and its context

Let us take an ambiguous word with the following morphological parsing:

$$stem_i + tag_{i,1} + tag_{i,2} + \cdots + tag_{i,L}$$

where $stem_i = (stem_{i,1}, stem_{i,2}, \ldots, stem_{i,K})$, a lemma K symbols long of the i[th] parsing option; each $tag_{i,j}$ is the j[th] tag (morphological feature) of the i[th] parsing option (which contains L of such tags). To construct the vector of the lemma, a bidirectional LSTM is used on top of each symbol of the lemma; for the vector of morphological features, we use a bidirectional LSTM over the tags. First, the bidirectional LSTM creates g_x representation of the input vector $x = (x_1, x_2, \ldots, x_T)$ by computing the direct $\vec{g}$ and the inverse $\overleftarrow{g}$ sequence, and combines the two sequences using the Rectified Linear Unit (ReLU).

$$\vec{g}_t = f(x_t, \vec{g}_{t-1})$$
$$\overleftarrow{g}_t = f(x_t, \overleftarrow{g}_{t+1})$$
$$g = ReLU(\vec{g}_T, \overleftarrow{g}_0)$$

where f(x, y) is a LSTM function with input values x and y.

Thus, the corresponding vector representations are constructed separately for the lemma and for the tag sequence (morphological features). Next, the resulting vectors are combined using the hyperbolic tangent:

$$r_i = \tanh(g_{stem_i} + g_{tag_i})$$

Next, r_i vectors are combined into R matrix, where each row belongs to a particular parse.

$$R = [r_1, r_2, \ldots, r_N]$$

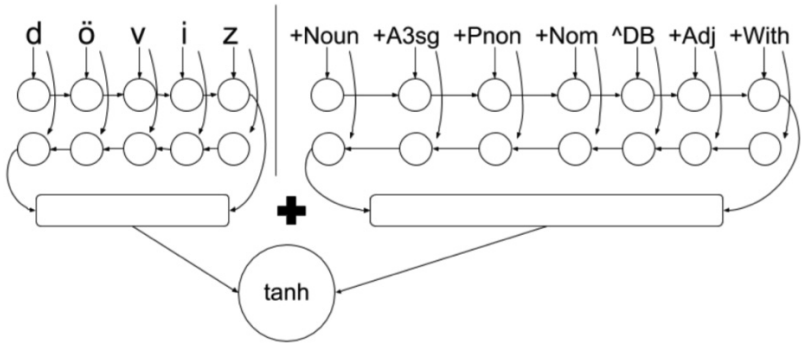


Fig. 1. LSTM neural network architecture for obtaining a vector representation of the morphological parse

One of the methods for constructing the context vector described in [5] is to use only the surface forms of the surrounding words (without morphological features). For this, the bidirectional LSTM model is used over each x_i word, constructing a separate vector for each word. Then for the left context, the vectors are assembled from right to left, and for the right context – from left to right (see Fig. 2.). After that, the vectors are combined using the hyperbolic tangent:

$$\vec{c}_t = f(x_t, \vec{c}_{t-1})$$

$$\overleftarrow{c}_t = f(x_t, \overleftarrow{c}_{t+1})$$

$$h_t = \tanh(\vec{c}_t, + \overleftarrow{c}_t)$$

Next, in order to perform the morphological disambiguation, the distribution of alternative probabilities is constructed – for this, *softmax* function from the product of h_t vector and R matrix is taken, and the most probable parse is selected as the correct one (according to the same formula as described in the previous section):

$$p(y_t = a|x) = softmax(R_{x_t} vo\ h_t)$$

Sometimes, surface forms of the surrounding words in the context are not enough for morphological disambiguation. Apart from these, it is necessary that all ambiguities in the surrounding words are resolved, i.e. data on the lemma and on all morphological features corresponding to the given context are needed. In such cases, the remedy is sequential disambiguation, when information about the

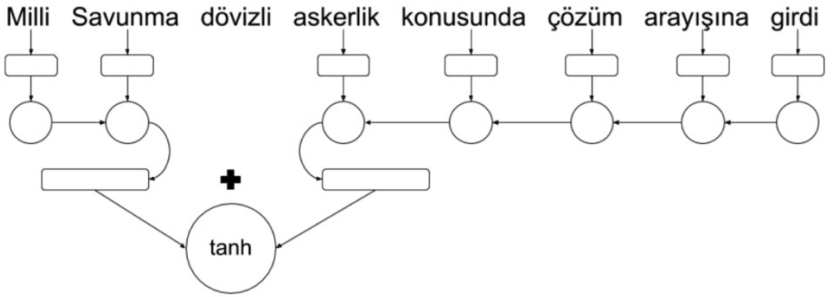


Fig. 2. Neural network architecture for obtaining a context vector.

allowed option is transmitted further, and the next case of ambiguity is resolved on its basis (in [5], this approach is defined as Left-to-Right).

In such cases, the LSTM model builds a vector based on the lemma and morphological features of the word from the context (if they are ambiguous, then the one in favor of which the disambiguation was made is selected) and thus m_t vector is calculated and then the disambiguation is performed:

$$m_t = f(r\,\mathrm{A}^t_i,\, m_{t-1})$$

where r_i^t is a vector from R_{x_t}, the parsing option selected at the previous disambiguation stage.

$$r_i^{t+1} = \tanh\left(g_{stem_i}^{t+1} + g_{tag_i}^{t+1} + m_t\right)$$

$$R_{x_{t+1}} = [r_1^{t+1}, r_2^{t+1}, \ldots, r_N^{t+1}]$$

$$p(y_t = a | x, y_1, y_2, \ldots, y_{t-1}) = softmax(R_{x_t} \times h_t)$$

$$\hat{y} = argmax_{\tilde{y} \in Y_X} \prod_{i=1}^{T} p(\tilde{y}_t | x, \tilde{y}_1, \tilde{y}_{2,\ldots,} \tilde{y}_{t-1})$$

4 Data preparation

At the initial stage of work, statistical data on the frequency of word forms with multivariate parses, presented in Table 2, were obtained from the text base of the National corpus of the Tatar language “Tugan tel” [3]. The morphological module implemented on the basis of the HFST toolkit is used for the morphological tagging of the corpus. [14].

Table 2. Distribution of morphological parsing options

Parsing options	Number	Share in the corpus
Total number of word forms with multivariate parses	5.650.820	25,75%
2 parses	4.282.108	19,51%
3 pares	1.045.392	4,76%
4 parses	296.547	1,35%
5 and more parses	26.773	0,12%
Total in the sample	21.940.452	100%

The total volume of the corpus at this stage was 21.940.452 tokens; the share of tokens with multivariate parses was 25.75%.

At the same time, the maximum length of the word form presented in the corpus consists of the stem and twelve grammatical affixes.

To carry out experiments with model training, it was necessary to have a morphologically disambiguated corpus. The part of socio-political sub-corpus of the National corpus of the Tatar language "Tugan tel" was used as training data. The sub-corpus statistics are given in Table 3.

Table 3. Statistics of the training and test samples on the socio-political corpus

	Training sample	Test sample
Number of contexts (sentences)	54.580	944
Number of tokens (including punctuation)	600.480	11.655
Number of multivariate parses	125.480 (21%)	2.527 (21%)
Number of unique word forms	29.953	2.788
Number of unique lemmas	7.117	1.226
Number of unique morphological forms	1.898	346

Manual morphological disambiguation of the socio-political sub-corpus was carried out by experts using a Web-based toolkit for morphological disambiguation in the corpus of the Tatar language [15].

Manual morphological disambiguation was organized in several stages.

At stage 1 selected texts from socio-political sub-corpus were automatically tagged using the morphological analyze. Then certain types of ambiguity were

automatically disambiguated where possible, as well as redundant and incorrect parses were removed.

At stage 2 annotators performed manual disambiguation using web-toolkit for morphological disambiguation in dialog mode. They selected the right parsing option based on the context.

At stage 3 main experts performed total manual review of the tagged texts disambiguated at stage 2. This double-checking helped us make sure that the tagging and disambiguation of the training data is correct.

As a result, for our experiments 56.524 morphologically disambiguated sentences were prepared.

5 Experiments and Evaluation

As one can see from Table 2, the tagged data sample was divided into a training sample and a test sample. LSTM models were trained only using the training set, and the test sample was used just for testing. Based on approach described in [5], we considered each sentence to be a minibatch for training. The objective function used for training was the total cross-entropy loss between the selected parse and the correct parse for every token in the sentence. Stochastic gradient descent and backpropagation were used to adjust the parameters for our model. All LSTMs in our models were trained with a single hidden layer. We used a hidden dimension size of 100 for the tag, stem, and surface form LSTMs and 200 for the context and previous parse LSTMs.

Tables 4, 5 provide an estimate of the accuracy of several indicators: lemma recognition, morpheme sequence recognition and disambiguation.

Table 4. Indicators of accuracy of recognition of lemmas and morpheme sequences

Indicators	LSTM NN
Lemma Recognition Accuracy	11299 / 11655 = 96.94%
Morpheme Sequence Recognition Accuracy	11127 / 11655 = 95.46%

Table 5 shows how the algorithm processes the different types of ambiguity according to the number of parsing options. As expected, the best result is for words with only two parsing options: 84.61%, when overall accuracy is 79.10%. In one hand, more variants increase complexity, in another hand, such words (see Table 2) do not have enough examples, so as a result, model lacks accuracy with them.

Table 5. The number of morphological parsing options and accuracy of disambiguation

Number of options	LSTM NN
n=2	1545 / 1826 = 84.61 %
n=3	268 / 424 = 63.21 %
n=4	141 / 192 = 73.44 %
n=5	7 / 9 = 77.78 %
n=6	37 / 72 = 51.39 %
n=7	0 / 2 = 0.00 %
n=8	0 / 1 = 0.00 %
Total	1999 / 2527 = 79.10%

The results of LSTM are virtually close to those of other disambiguation methods. The main benefit of the proposed method is that the model can be trained taking into account the size and peculiarities of the context. So the highest accuracy rate of the morphological disambiguation in the corpus of the Tatar language was achieved with the construction of vectors taking into account all the words in the sentence as the surrounding context. In addition, the vector of the surrounding context was expanded using morphological features.

6 Conclusions

This paper presents the results of work on morphological disambiguation of the Tatar language using the neural network model based on the LSTM architecture. Given the limited set of corpus data for training, the results of experiments showed a fairly good level of accuracy for morphological disambiguation, 79.10%. We believe that the lower accuracy of the neural network model is primarily related to the amount of training data, since systems with neural networks are not sufficiently effective when training on a limited set of data.

At the same time, the obtained results can be effectively used in creating a morphologically disambiguated “golden” sub-corpus, significantly reducing the number of multivariate parses requiring manual morphological disambiguation.

References

1. Gataullin, R.R.: Analiticheskij obzor metodov razresheniya morfologicheskoj mnogo-znachnosti [Analytical review of the methods of morphological disambiguation]. Rossijskij nauchnyj elektronnyj zhurnal (Elektronnye biblioteki) 19(2), 98-114 (2016).
2. Zin’kina, Yu.V., Pyatkin, N.V., Nevzorova, O.A.: Razreshenie funkcional’noj

omonimii v russkom yazyke na osnove kontekstnyh pravil [Context rule based functional disambiguation in Russian]. In: Proceedings of the International Conference "Dialog-2005", pp. 198–202. Nauka, Moscow (2005).
3. Gataullin, R., Khakimov, B., Suleymanov, D., Gilmullin, R.: Context-Based Rules for Grammatical Disambiguation in the Tatar Language. In: Nguen, N.T. et al. (eds). ICCCI 2017, Part II, LNAI, vol.10449, pp. 529-537 (2017).
4. Sak, H., Gongur, T., Saraclar. M.: Morphological disambiguation of Turkish text with perceptron algorithm. In: Computational Linguistics and Intelligent Text Processing, 8th International Conference CICLing, Mexico City, Mexico, February 2007, pp. 107–118. Mexico (2007).
5. Qinlan Shen, Daniel Clothiaux, Emily Tagtow, Patrick Littell, Chris Dyer.: The Role of Context in Neural Morphological Disambiguation. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 181–191. Osaka (2016) http://aclweb.org/anthology/C16-1018, last accessed 2018/10/25.
6. Orosz, G. and Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of Recent Advances in Natural Language Processing, pp. 539–545. (2013) http://aclweb.org/anthology//R/R13/R13-1071.pdf, last accessed 2018/10/25.
7. Yuret, D., Ture, F.: Learning morphological disambiguation rules for Turkish. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pp. 328–334. New York (2006).
8. Gilmullin, R.A., Gataullin, R.R.: Razreshenie morfologicheskoj mnogoznachnosti tekstov na tatarskom yazyke na osnove instrumentariya PurePos [Morphological disambiguation of Tatar texts using PurePos]. In: Proceedings of the V International Conference on Turkic Languages Processing «TurkLang 2017», pp. 30-37. Fen, Kazan (2017).
9. Berment, V.: Me´thodes pour informatiser des langues et des groups de langues peu dote´es, Ph.D. Thesis, J. Fourier University, Grenoble (2004).
10. Krauwer, S.: The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In: Proceedings of International Workshop on Speech and Computers - SPEECOM, pp. 8–15. Moscow (2003).
11. Hochreiter, S.: Long short-term memory. Neural Computation 9(8), 1735–1780 (1997).
12. Gers, F.A.: Learning to Forget: Continual Prediction with LSTM. Neural Computation 12(10), 2451–2471 (2000).
13. Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.): Ethnologue: Languages of the World, http://www.ethnologue.com last accessed 2018/10/25.
14. Gilmullin, R., Gataullin, R.: Morphological Analysis System of the Tatar Language. In: Nguen, N.T. et al. (eds). ICCCI 2017, Part II, LNAI, vol.10449, pp. 519-528 (2017).
15. Gataullin, R.R.: Web-instrumentarij dlya snyatiya morfologicheskoj mnogoznachnosti v tekstovom korpuse tatarskogo yazyka [Web-based toolkit for morphological disambiguation in the corpus of Tatar texts]. In: Proceedings of the V International Conference on Preservation and Development of Native Languages in a Multinational State, pp. 71-73. Otechestvo, Kazan (2014).

Extended Language Modeling experiments for Kazakh

Bagdat Myrzakhmetov[1,2] and Zhanibek Kozhirbayev[1]

[1] National Laboratory Astana, Nazarbayev University, Astana, 010000, Kazakhstan
[2] Nazarbayev University, School of Science and Technology, Astana, 010000, Kazakhstan
bagdat.myrzakhmetov@nu.edu.kz , zhanibek.kozhirbayev@nu.edu.kz

Abstract. In this article we present dataset for the Kazakh language for the language modeling. It is an analogue of the Penn Treebank dataset for the Kazakh language as we followed all instructions to create it. The main source for our dataset is articles on the web-pages which were primarily written in Kazakh since there are many new articles translated into Kazakh in Kazakhstan. The dataset is publicly available for research purposes[1]. Several experiments were conducted with this dataset. Together with the traditional n-gram models, we created neural network models for the word-based language model (LM). The latter model on the basis of large parameterized long short-term memory (LSTM) shows the best performance. Since the Kazakh language is considered as an agglutinative language and it might have high out-of-vocabulary (OOV) rate on unseen datasets, we also carried on morph-based LM. With regard to experimental results, sub-word based LM is fitted well for Kazakh in both n-gram and neural net models compare to word-based LM.

Keywords: Language Modeling, Kazakh language, n-gram, neural language models, morph-based models.

1 Introduction

The main task of the language model is to determine whether the particular sequence of words is appropriate or not in some context, determining whether the sequence is accepted or discarded. It is used in various areas such as speech rec-

[1] https://github.com/Baghdat/LSTM-LM/tree/master/data/

ognition, machine translation, handwriting recognition [1], spelling correction [2], augmentative communication [3] and Natural Language Processing tasks (part-of-speech tagging, natural language generation, word similarity, machine translation) [4, 5, 6]. Strict rules may be required depending on the task, in which case language models are created by humans and hand constructed networks are used. However, development of the rule-based approaches is difficult and it even requires costly human efforts if large vocabularies are involved. Also usefulness of this approach is limited: in most cases (especially when a large vocabulary used) rules are inflexible and human mostly produces the ungrammatical sequences of words during the speech. One thing, as [7] states, in most cases the task of language modeling is "to predict how likely the sequence of words is", not to reject or accept as in rule-based language modeling. For that reason, statistical probabilistic language models were developed.

A large number of word sequences are required to create the language models. Therefore the language model should be able to assign probabilities not only for small amounts of words, but also for the whole sentence. Nowadays it's possible to create large and readable text corpora consisting of millions of words, and language models can be created by using this corpus.

In this work, we first created the datasets for the language modeling experiments. We built an analogy of the Penn Treebank corpus for the Kazakh language and to do so we followed all preprocessing steps and the corpus sizes. The Penn Treebank (PTB) Corpus [8] is widely used dataset in language modeling tasks in English. The PTB dataset originally contains one million words from the Wall Street Journal, small portion of ATIS-3 material and tagged Brown corpus. Then [9] preprocessed this corpus, divided into training, validation and test sets and restricted the vocabulary size to 10k words. From then, this version of PTB corpus is widely in language modeling experiments for all state of the art language modeling experiments. We made our dataset publicly available for any research purposes. Since there are not so many open source corpora in Kazakh, we hope that this dataset can be useful in the research community.

Various language modeling experiments were performed with our dataset. We first tried traditional n-gram based statistical models, after that performed state-of-the-art Neural Network based language modeling experiments. Neural Network experiments were conducted by using the LSTM [10] cells. LSTM based neural network with large parameters showed the best result. We evaluated our language modeling experiments with the perplexity score, which is a widely used metric to evaluate language models intrinsically. As the Kazakh language is agglutinative language, word based language models might have high portion of out of vocabulary (OOV) words on unseen data. For this reason, we also performed morpheme-based language modeling experiments. Sub-word based

language model is fitted well for Kazakh in both n-gram and neural net models compare to word-based language models.

2 Data preparation

We collected the datasets from the websites by using our manual Python scripts, which uses BeautifulSoup and Request libraries in Python. These collected datasets were parsed with our scripts on the basis of the HTML structure. The datasets were crawled from 4 web-pages, whose articles originally written in Kazakh: `egemen.kz`, `zhasalash.kz`, `anatili.kazgazeta.kz` and `baq.kz`. These web-pages mainly contain news articles, historical and literature texts. There are many official web-pages in Kazakhstan which belong to state bodies and other quasi-governmental establishments where texts in Kazakh could be collected. However, in many cases, these web-pages provide the articles, which were translated from the Russian language. In these web-pages, the news articles at the beginning will be written in Russian, only then, these articles translated into Kazakh. These kind of datasets might not well show the inside nature of the Kazakh language, as during the translation, the structure of the sentences and the use of words changes. We barely see the resistant phraseological units of Kazakh in these translated articles, instead we might see the translated version of the phraseological texts in other language. [11] studied original and translated texts in Machine translation, and found out that original texts might be significantly differing from the original texts. For this reason, we excluded the web-pages which might have translation texts. We choose the web-pages whose texts originally written in Kazakh. The statistics of datasets is given in Table 1.

Table 1. Statistics of the dataset: train, validation and test sets shown separately for each source.

Sources	# of documents	# of sentences	# of words
egemen.kz	950/80/71	21751/1551/1839	306415/22452/26790
zhasalash.kz	1126/83 /95	8663/694/751	102767/8188/9130
anatili.kazgazeta.kz	438/32/37	23668/1872/2138	311590/23703/27936
baq.kz	752/72/74	13899/1082/1190	168062/13251/14915
Overall	3266/267/277	67981/5199/5918	886872/67567/78742

After collection of the datasets, we preprocessed the datasets by following [9]. First, all collected datasets were tokenized using Moses [12] script. We added non-breaking prefixes for Kazakh in Moses, so as not to split the abbreviations.

Next preprocessing steps involved: lowercasing, normalization of punctuations. After normalization of the punctuations, we removed all punctuation signs. All digits were replaced by a special sign “N”. We removed all sentences whose length is shorter than 4 and longer than 80 words and also duplicate sentences. After these operations, we restricted the vocabulary size with 10000: we found the most frequent 10000 words and then replaced all words with ‘<unk>’, which are not in the list of the most frequent words.

After preprocessing of the datasets, we divided our datasets into training, validation and testing sets. We tried to follow the size of the Penn Treebank corpus. Since our datasets were built from the four sources, we tried to split all sources in the same proportion into training, validation and test sets. Since, the contents in each source might differ (for example, in egemen.kz there are mostly official news, on the other hand anatili.kazgazeta.kz contains mainly historic, literature articles), we avoid having one source as training and others only for testing or validation. For this reason, we split each source with equal portions. Our datasets divided into training, validation and test sets on the document level. The statistics about training, validation and test sets is given in Table 2. Note, overall sentence and word numbers might not be the sum of all columns, because we exclude the repeated sentences. To compare the size, at the end, we provide the statistics of the Penn Treebank corpus.

Table 2. Statistics about the training, validation and test sets.

Sources	Train set	Validation set	Test set
egemen.kz	306415	22452	26790
zhasalash.kz	102767	8188	9130
anatili.kazgazeta.kz	311590	23703	27936
baq.kz	168062	13251	14915
Overall	886872	67567	78742
Penn Tree Bank dataset	887521	70390	78669

3 N-gram based models

The main idea behind the language modeling is to predict hypothesized word sequences in the sentence with the probabilistic model. “N-gram models predict the next word from the previous N-1 words” and it is an N-token sequence of words, [13] for example, if we say two-gram model (or more often it is called a bigram model) it is two-word sequence such as “Please do”, “do your”, “your homework” and three gram model consists of the three-word sequences and so

on. As [13] states, in n-gram model, the model computes the following word from the preceding. The N-gram idea can be formulated as: given the pervious word sequence and find the probability of the next words. During the computing of probabilities of the word sequences it's important to define the boundaries (punctuation marks such as period, comma, column or starting of the new sentence from the new line) in order to prevent the search from being computationally unmanageable.

Formulated mathematically, the goal of a language model is to find the probability of word sequences, $P(w_1, \ldots, w_n)$, and it can be estimated by the chain rule of a probability theory:

$$P(w_1, \ldots, w_n) = P(w_1) \times P(w_2|w_1) \times \ldots \times P(w_n|w_1, \ldots, w_{n-1}) \quad (1)$$

There is a notion about history, for example, in the case $P(w_4|w_1, w_2, w_3)$, (w_1, w_2, w_3) considered as the history. This probability is found based on frequency.

We can write the formula for all cases bigram and trigram models as:

$$P(w_i|w_1 \ldots w_{i-1}) \approx P(w_i|w_{i-1}) \quad (2)$$

$$P(w_i|w_1 \ldots w_{i-1}) \approx P(w_i|w_{i-2}w_{i-1}) \quad (3)$$

This assumption helps to reduce the computation and allows probabilities to be estimated for a large corpus. Also the assumption probability of the word which depends on the previous n words (or previous 3 words for a trigram) is called a **Markov assumption**. This Markov model [14] assumes that it is possible to predict the probability of some future cases without looking deeply into the past.

By using a Markov assumption, we can find the probability of the sequence of words by the following formula:

$$P(w_1, \ldots, w_n) = \prod P(w_i|w_1 \ldots w_{i-1}) \approx \prod w(w_i|w_{i-1}) \quad (4)$$

for bigram model and for trigram:

$$\approx \prod (w_i|w_{i-2}w_{i-1}) \quad (5)$$

Up to recently, n-gram language models widely used in all language modeling experiments. In Kazakh, n-gram based language models still used in Speech Processing [15] and Machine translation [16] tasks. We trained n-gram models with the SRILM toolkit [17] with adding 0 smoothing technique. For our dataset, using of the modified Kneser-Ney [18] or Katz backoff [19] algorithms

showed poor results, (543.63 on the test set), as there are many infrequent words replaced by '<unk>' sign, and only high gram models might work well. Adding 0 smoothing technique showed best performance for n-gram models. The results are given in Table 3.

4 Neural LSTM based models

In this experiment, we performed Neural LSTM-based language models. There are many types of neural architectures, which also applied successfully for the language modeling tasks. Starting from the work of [20] there are many Recurrent Neural Architectures proposed. With Recurrent Neural Networks, it's possible to model the word sequences, as the recurrence allows to remember the previous word history. Recurrent Neural Network can directly model the original conditional probabilities:

$$P(w_1, \ldots, w_n) = \prod P(w_i | w_1 \ldots w_{i-1}) \quad (6)$$

To model the sequences, f function constructed via recursion, initial condition is given by $h_0 = 0$ and the recursion will be $h_t = f(x_t, h_{t-1})$. Here, h_t is called hidden state or memory and it memorizes the history from x_1 up to x_{t-1}. Then, the output function is defined by combination of h_t function:

$$P(w_1, \ldots, w_n) = g_w(h_t) \quad (7)$$

f can be any nonlinear function such as *tanh*, *ReLU* and *g* can be a *softmax* function.

In our work, we followed [21] who presented a simple regularization technique for Recurrent Neural Networks (RNNs) with LSTM [10] units. [22] proposed dropout technique for regularizing the neural networks, but this technique does not work well with RNNs. This regularizing technique is tent to have overfitting in many tasks. [21] showed that the correctly applied dropout technique to LSTMs might substantially reduce the overfitting in various tasks. They tested their dropout techniques on language modeling, speech recognition, machine translation and image caption generation tasks.

In general, LSTM gates' equations given as follow:

$$f_t = \sigma(W_f[C_{t-1}, h_{t-1}, x_t] + b_f) \quad (8)$$

$$i_t = \sigma(W_i[C_{t-1}, h_{t-1}, x_t] + b_i) \quad (9)$$

$$o_t = \sigma(W_o[C_t, h_{t-1}, x_t] + b_o) \quad (10)$$

$$g_t = tanh(W_g[C_t, h_{t-1}, x_t] + b_g) \quad (11)$$

Then the state values computed by using the above gates:

$$c^l_t = f \odot c^l_{t-1} + i \odot g \quad (12)$$

$$h^l_t = o \odot tanh(c^l_t) \quad (13)$$

The dropout method by [21] can be described as follows: if there is a dropout operator, then it forces the intermediate computation to be more robustly, as the dropout operator corrupts the information carried by the units. On the other hand, in order not to erase all the information from the units, the units remember events that occurred many time steps in the past.

We also implement our[1] LSTM based Neural Network models using TensorFlow [23]. We trained regularized LSTMs of three sizes: the small LSTM, medium LSTM and large LSTM. Small sized model has two layers and unrolled for 20 steps. Medium and large LSTMs have two layers and are unrolled for 35 steps. Hidden size differs in three models: 200, 650 and 1500 for small, medium and large models respectively. We initialize the hidden states to zero. We then use the final hidden states of the current minibatch as the initial hidden state of the subsequent minibatch.

Our experiments showed that the LSTM based neural language modeling outperforms the n-gram based models. Large and Medium LSTM models shows better results than the n-gram add 0 smoothing method (Note, for n-gram Kneser-Ney discounting method we got poor results). Our experiments show that the using of the Neural based language models have better performance for Kazakh. The results are given in Table 3.

Table 3. Word-based language modeling results.

	n-gram	Neural LM		
		small	medium	large
Train ppl	93.81	68.522	67.741	**63.185**
Validation ppl	129.6537	143.871	118.875	**113.944**
Test ppl	123.7189	144.939	118.783	**115.491**

5 Sub-word based language models

In the last section, we experimented with the sub-word based language models. The Kazakh language as other Turkic languages is an agglutinative language,

[1] https://github.com/Baghdat/LSTM-LM

the word forms can be obtained by adding the prefixes. This agglutinative nature may lead on having the high degree of the out-of-vocabulary (OOV) words on unseen data. To solve this problem, depending on the characteristics of individual languages, different language model units were proposed. [24] studied different word representations, such as morphemes, word segmentation based on the Byte Pair Encoding (BPE), characters and character trigrams. Byte Pair Encoding, proposed by [25], can effectively handle rare words in Neural Machine Translation and it iteratively replaces the frequent pairs of characters with a single unused character. Their experiments showed that for fusional languages (Russian, Czech) and for agglutinative languages (Finnish, Turkish) character trigram models perform best. Also, [26] considered syllables as the unit of the language models and tested with different representational models (LSTM, CNN, summation). As they stated, syllable-aware language models fail to outperform character-aware ones, but usage of syllabification can increase the training time and reduce the number of parameters compared to the character-aware language models.

By considering these facts, in this section we experimented with the sub-word based models. Morfessor [27] is a widely tool to split the datasets into morpheme-like units. It used successfully in many agglutinative languages (Finnish, Turkish, Estonian). As for now, there is no syllabification tool for Kazakh, we also used Morfessor tool to split our datasets into morpheme like units.

After splitting the datasets, we performed language modeling experiments on morpheme like units. The results are given in Table 4. By looking at the results, we can say that splitting the words into morpheme-like units benefits in terms of OOV and perplexity in both n-gram and neural net based models.

Table 4. Morph-based language modeling results.

	n-gram	Neural LM		
		small	medium	large
Train ppl	32.39255	19.599	24.999	25.880
Validation ppl	44.11561	50.904	41.896	40.876
Test ppl	44.39559	47.854	38.180	**37.556**

6 Conclusion

In this work we created analogy of the Penn TreeBank corpus for the Kazakh language. To create the corpus, we followed all instructions for preprocessing

and the size of the training, validation and test sets. This dataset is publicly available for the research purposes. We conducted language modeling experiments on this dataset by using the traditional n-gram and LSTM based neural networks. We also explored the sub-word units for the language modeling experiments for Kazakh. Our experiments showed that neural based models outperform the n-gram based models and splitting the words into morpheme-like units has advantage compared to the word based models. In future, we are going to create the hyphenation tool for the Kazakh language, as Morfessor's morpheme-like units are data-driven and sometimes there are incorrect morpheme-like units.

Acknowledgement

This work has been funded by the Nazarbayev University under the research grant No129-2017/022-2017 and by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan under the research grant AP05134272.

References

1. Russell S. and Norvig P. Artificial Intelligence: A Modern Approach (2nd Ed.). Pretice Hall. 2002.
2. Kukich K. Techniques for automatically correcting words in text. ACM Computing Surveys.1992. 24(4), pp. 377-439.
3. Newell A., Langer S. and Hickey M. The role of natural language processing in alter-native and augmentative communication. Natural Language Engineering. 1998. 4(1). pp. 1-16.
4. Church K.W. A stochastic parts program and noun phrase parser for unrestricted text. In Proceedings of the Second Conference on Applied Natural Language Processing. 1988. pp. 136–143.
5. Brown P.F., Cocke J., DellaPietra S.A., DellaPietra V.J., Jelinek F., Lafferty J.D., Mercer R.L., and Roossin P.S. A statistical approach to machine translation. Computational Linguistics. 1990. 16(2). pp. 79–85.
6. Hull J.J. Combining syntactic knowledge and visual text recognition: A hidden Markov model for part of speech tagging in a word recognition algorithm. In AAAI Symposium: Probabilistic Approaches to Natural Language. 1992. pp. 77–83.
7. Whittaker E. W. D. Statistical Language Modelling for Automatic Speech Recognition of Russian and English. PhD thesis, Cambridge University, Cambridge. 2000.
8. Marcus M.P., Marcinkiewicz M.A. and Santorini B. Building a large annotated corpus of English: The penn Treebank. Computational linguistics. 1993. 19(2). pp. 313–330.

9. Mikolov T., Kombrink S., Burget L., Černocký J. and Khudanpur S. Extensions of recurrent neural network language model. In Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. 2011. pp. 5528-5531. IEEE.
10. Hochreiter S. and Schmidhuber J̈. Long short-term memory. Neural computation. 1997. 9(8). pp. 1735–1780.
11. Lembersky G., Ordan N. and Wintner S. Language models for machine translation: Original vs. translated texts. Computational Linguistics. 2012. 38(4). pp. 799-825.
12. Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R. and Dyer C. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. Association for Computational Linguistics. 2007. pp. 177-180.
13. Jurafsky D. and Martin J. H. Speech and Language Processing (2nd Ed.). Pretice Hall. 2009.
14. Markov A.A. Primer statisticheskogo issledovaniya nad tekstom "Evgeniya Onegina", illyustriruyushchij svyaz' ispytanij v tsep'. [Example of a statistical investigation illustrating the transitions in the chain for the "Evgenii Onegin" text.]. Izvestiya Akademii Nauk. 1913. pp. 153-162.
15. Kozhirbayev Zh, Karabalayeva M. and Yessenbayev Zh. Spoken term detection for Kazakh language. In Proceedings of the 4-th International Conference on Computer Processing of Turkic Languages "TurkLang 2016". 2016. pp. 47-52.
16. Myrzakhmetov B. and Makazhanov A. Initial Experiments on Russian to Kazakh SMT. Research in Computing Science. 2017. vol. 117. pp. 153–160.
17. Stolcke, A. SRILM – an extensible language modeling toolkit. In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP). 2002. pp. 901–904. URL: http://www.speech.sri.com/ projects/srilm/.
18. Kneser R. and Ney H. Improved backing-off for m-gram language modeling. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. 1995. vol. 1. pp. 181-184.
19. Katz S. M. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustics, Speech and Signal Processing. 1987. 35(3). pp. 400-401.
20. Bengio Y., Ducharme R., Vincent P. & Jauvin C. A neural probabilistic language model. Journal of machine learning research. 2003. pp. 1137-1155.
21. Zaremba W., Sutskever I. and Vinyals O. Recurrent neural network regularization. arXiv preprint arXiv:1409.2329. 2014.
22. Srivastava N., Hinton G., Krizhevsky A., Sutskever I. & Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research. 2014. 15(1). pp. 1929-1958.
23. Abadi M., Barham P., Chen J., Chen Zh., Davis A., Dean J., Devin M., Ghemawat S., Irving G., Isard M. and Kudlur M. Tensorflow: a system for large-scale machine learning. In Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation. USENIX Association. 2016. pp. 265–283.

24. Vania, C., & Lopez, A. From Characters to Words to in Between: Do We Capture Morphology? In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. Volume 1: Long Papers. Vol. 1, pp. 2016-2027.
25. Sennrich, R., Haddow, B., & Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016. Volume 1: Long Papers. Vol. 1, pp. 1715-1725.
26. Assylbekov Z., Takhanov, R., Myrzakhmetov, B., & Washington, J. N. Syllable-aware Neural Language Models: A Failure to Beat Character-aware Ones. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017. pp. 1866-1872.
27. Smit P., Virpioja S., Grönroos S. A. & Kurimo M. Morfessor 2.0: Toolkit for statistical morphological segmentation. In The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014. Aalto University.

Roles Contradictions Play in Logical Models of Metaphors and Presuppositions

Boris Kulik[1] and Alexander Fridman[2]

[1] Institute of Problems in Mechanical Engineering of RAS, St. Petersburg, Russia
[2] Institute for Informatics and Mathematical Modelling, Kola Science Centre of RAS, Apatity, Russia
ba-kulik@yandex.ru, fridman@iimm.ru

Abstract. The article considers new approaches to logical analysis of metaphors and presuppositions. Arguments are given in favor of the necessity for a contradiction to be present in metaphors. It is established that logical model of presuppositions also contains a contradiction, but, unlike in metaphors, here it can be eliminated without distortion the meaning of the phrase.

Keywords: metaphor, presupposition, logical model, contradiction.

1 Introduction

In natural languages, including scientific texts and journalism, and especially in subsets of natural languages used in computers, any contradiction is considered to be an undesirable component, which should be avoided as much as possible. At the same time, contradictions are inevitable and, as a rule, they stimulate criticism and development of our knowledge. However, situations are widely known where contradictions are considered not disadvantages, but advantages of a language. These include metaphors, which will be shown to not exist without a hidden or obvious contradiction. In addition, we will put forward arguments in favor of the necessity for a contradiction to be contained in the logical model of presupposition. Conversely to the metaphor, in this case it can be eliminated without distorting the meaning of the text.

Let us clarify what is meant here by a contradiction. In formal logic, a contradiction is defined as an identically false logical formula in which any substitutions for any interpretation are false. For example, formulas $A \wedge \neg A$ and $(A \supset B) \wedge (A \supset \neg B) \wedge A$ contain contradictions. In natural reasoning, the concept

of contradiction is broader: a situation is considered a contradictory one when an object is supposed to exist within the situation and this object has incompatible properties. Consider an example when the following premises are given:

1) All my friends are braggarts.

2) All my friends are not rowdies.

3) All braggarts are rowdies.

To analyze this reasoning, we can use the means of propositional calculus [1] or partially ordered sets [2], but we will use a relatively simple system of logical inference described in [3] and based on *QC*-structures.

A ***QC*-structure** (abbreviated from quasi-complement) is a partially ordered set (poset) that has the *smallest* (**0**) and the *largest* (**1**) elements and a quasi-complement operation with the following properties:

(i) for any element A of a poset, there exists or can be computed a single element $\overline{A}$ called the ***quasi-complement*** of A;

(ii) for any element A, the equality $\overline{\overline{A}} = A$ is satisfied;

(iii) for any two elements A and B, if $A \le B$, the contraposition $\overline{B} \le \overline{A}$ is correct.

This mathematical system completely describes properties of all types of partially ordered sets, multisets, and fuzzy sets. This system is proved to not comply with the law of the excluded middle, which is typical, in particular, for fuzzy sets and multisets. But if we extend the axioms of quasi-complement with the property:

(iv) for any element A, the relation $A \le \overline{A}$ is admissible only for the case when $A = \mathbf{0}$ and $\overline{A} = \mathbf{1}$,

we obtain a poset that has all properties of the inclusion relation in algebra of sets. Such kind of *QC*-structures is called ***Euler's logical structure*** (the name is due to the fact that these structures correspond to the properties of Euler's circles (or diagrams)) abbreviated as ***E*-structure**. In *E*-structures, the order relation is usually denoted by the symbol "$\subseteq$".

Universal affirmative propositions of the type "All A are B" or "The property B is inherent to the object A" can be represented as set inclusions: $A \subseteq B$. Universal negative propositions of the type "All A are not B" we model as $A \subseteq \overline{B}$. Unlike Aristotelian syllogistics, the reasoning system [3], in accordance with a more accurate modern conception of logical deduction, admits arranging premises in an arbitrary order (in syllogistics, conclusions change in some cases, when the order of premises changes), and the first premise of a statement may be negative, which is not recommended in syllogistics.

If there are no particular judgments in a system of premises [3], it is sufficient to use only two laws of algebra of sets as rules of inference, namely:

1) *contraposition* ($A \subseteq B$ is equipotent to $\overline{B} \subseteq \overline{A}$) and
2) *transitivity* ($A \subseteq B$ and $B \subseteq C$ infer $A \subseteq C$).

To reduce complexity of calculations when working with a large number of initial premises of the type $X \subseteq Y$, it is recommended to apply the contraposition law for all premises of reasoning first, after which all other consequences can be obtained by using the law of transitivity.

A contradiction in the broad sense (a paradox collision [3]) is stated here in the case when an inference results in at least one premise of the type $X \subseteq \overline{X}$.

Denote F for my friends, B as braggarts, R as rowdies. Then the premises can be written as logical formulas: 1) $F \subseteq B$; 2) $F \subseteq \overline{R}$; 3) $B \subseteq R$. Consider the corollaries. By the contraposition rule, $F \subseteq \overline{R}$ derives $R \subseteq \overline{F}$, and $B \subseteq R$ infers $\overline{R} \subseteq \overline{B}$. According to the transitivity rule, $F \subseteq B$ and $B \subseteq R$ yield $F \subseteq R$, $F \subseteq \overline{R}$ and $\overline{R} \subseteq \overline{B}$ infer $F \subseteq \overline{B}$, while $F \subseteq R$ and $R \subseteq \overline{F}$ deduce $F \subseteq \overline{F}$.

If we translate the resulting consequences into a natural language, we will see that “my friends” have opposite properties: they are “braggarts” and “not braggarts”, “rowdies” and “not rowdies”, and eventually it turns out that “my friends” are “not my friends”.

If we use the language of propositional calculus for this reasoning system, the inclusion $X \subseteq Y$ shall be replaced with the implication $X \supset Y$, complement of the set $\overline{X}$ is equal to the negation $\neg X$, and the totality of premises shall be united by using the symbol of conjunction ($\wedge$).

Then the enumerated premises can be expressed in the form of the logical formula: $(F \supset B) \wedge (F \supset \neg R) \wedge (B \supset R)$. Analysis that is not included here shows that this formula is not identically false. Hence, there is no formal contradiction, but the object “my friends” does not exist here. In natural reasoning, such a broad-sense contradiction denotes non-existence of a presumably existing object. In [3], it is called a paradox collision. It would be a good idea to use this term in order not to confuse the broad-sense contradiction with a formal one, but we propose to keep the more common name in this article. We define this case as a paradox since the object F is assumed to be true (that is, I have friends), while the premises yield that this object is false. This situation can be reduced to a formal contradiction, if we add what is meant (in this example, the formula F), to the initial premises. Then, it is easy to prove that the formula

$$F \wedge (F \supset B) \wedge (F \supset \neg R) \wedge (B \supset R),$$

which includes the implicitly expressed premise, is formally contradictory. Indeed, as proved above, the sub-formula $(F \supset B) \wedge (F \supset \neg R) \wedge (B \supset R)$ infers $\neg F$, and the formula $F \wedge \neg F$ is contradictory.

The simplest logical model of a contradiction of the given type is defined by the formula $A \wedge (A \supset B) \wedge (A \supset \neg B)$, which can be expressed by the following

phrase: "The object A is true; at the same time it has the property B and does not have the property B". This formula is easy to prove to be equivalent to the formula $\neg A$, but this does not mean that the given formula is identically false, since it contains a dummy variable B, whose value can be either true or false. In general case, we will consider a reasoning ***contradictory*** (in the broad sense), if its formalization and logical analysis reveal a variable X that denotes a presumably existing object and assumes the false value only.

2 Logical Model of Metaphors

A metaphor is a word (in general, an expression) that is intentionally used in the text instead of another (replaced) word (expression) based on some incomplete coincidence of meanings of these words (expressions). Such incomplete coincidence of meanings in the definition of a metaphor is essential; otherwise it is difficult to distinguish a metaphor from a synonym. Sometimes a metaphor is defined as an action. For example, "Metaphor is the transfer of a name from one subject or phenomenon to another based on their similarity in some respect." As O. N. Laguta [4] noticed, definitions of this kind use a metaphor (the word "transfer" is a metaphor). Note that usage of metaphors in any science is more likely a rule than an exception (for instance, locutions like "the effect of gossip" in chemical reactions, "black hole", "solar corona", "vertebral column", "computer virus", "lattices" in mathematics, etc.).

The concept of metaphor was known even in ancient Greece. Here is the Aristotle's definition: "Metaphor is a transfer of a word with an altered meaning from a genus to a species, from a species to a genus, or from a species to another species, or in the form of proportion." Metaphor is considered in numerous scientific works throughout the course of human history, beginning with antiquity. At present, the growth of research interest to the metaphor is associated with formation of cognitive science [4 – 7].

Interest to the metaphor becomes more intense and rapidly widens, capturing different fields of knowledge, namely philosophy, logic, psychology, psychoanalysis, hermeneutics, literary criticism, philology, theory of fine arts, semiotics, rhetoric, linguistic philosophy, and various schools of linguistics [5]. Due to this increased interest, a new science has emerged, whose name is "metaphorology" [4].

Let us consider some logical models of the metaphor. In [4], a logical model based on the "deviatological approach" is considered, where the metaphor manifests itself as some logical anomaly. Within this approach, a deviation from logical norms is detected when a metaphor is a convoluted deduction

(enthymeme), i.e., an inference with a missed premise. As an example, in [4] the metaphor "Admiralty Needle" (in the quoted text, it is "the Needle of the Admiralty") from Pushkin's poem "The Bronze Horseman" is used. Obviously, the "needle" replaces the word "spire" in this case. The following inference is proposed.

The minor premise:

This spire (architectural element) (S) is very long in relation to its own diameter, straight, with a point (M).

The major premise:

[Everyone knows that] some (tools) are long in relation to their own diameter, straight, with a point (M); they are needles (P).

Conclusion: The spire (S) is a needle (P).

Note that in this example, "deviation" is not only a "convolution" of the inference. According to the rules of syllogistics and formal logic, the conclusion "A spire is a needle" cannot be deduced from the initial premises, even if you do not take into account the bracketed differences ("architectural element" and "tool"). The reason is that the major premise is formulated as a private assertion (it contains the word "some"), so the syllogism turns out to be wrong, and the given conclusion is nondeductive.

In order to correctly formulate a logical model of the metaphor, it is necessary to recognize that such a model necessarily contains a contradiction. Traditionally, it is considered that only some properties of an object or a phenomenon, denoted by a metaphor, coincide with properties of an object denoted by a replaced word. At the same time, in the numerous definitions of the metaphor, the differences in meanings of the word-metaphor and the replaced word are not sufficiently emphasized; just this feature of the metaphor determines many of its remarkable attributes and, moreover, distinguishes it from another linguistic phenomenon – synonymy.

Consider the same example. We designate S as a spire, M as a very long object with respect to its own diameter, straight, with a point, P denotes a needle, A means an architectural element, and T is a tool.

Let us formulate the correct premises taking into account the above distinction. Then we obtain:

$$S \supset M;\ P \supset M;\ S \supset A;\ P \supset T;\ A \supset \neg T.$$

The last premise affirms that properties "architectural element" and "tool" are incompatible. To analyze these premises, we use the inference system [3] again. Analysis shows that there are no contradictions in the totality of the above premises.

Let us construct some corollaries. By the contraposition rule, the judgment $P \supset T$ yields $\neg T \supset \neg P$, and the statements $S \supset A$, $A \supset \neg T$ and $\neg T \supset \neg P$ infer $S \supset \neg P$ by the transitivity rule (i.e., a spire is not a needle). Hence, it is clearly impossible to deduce the proposition "a spire is a needle" as a consequence of these premises. And if we add the statement $S \supset P$ (that is, the proposition that defines the metaphor) to the system of premises, we will obtain the expression $\neg S$ as one of the consequences (i.e., the logical variable "spire" takes the value "false"). To obtain a formal contradiction, it suffices to add the formula S to the premises, which means that the replacing word "spire" is true.

If we use the means of mathematical logic, then this and any other metaphor is restored and reproduced by the logical formula

$$S \wedge (S \supset M) \wedge (P \supset M) \wedge (S \supset A) \wedge (P \supset T) \wedge (A \supset \neg T) \wedge (S \supset P), \quad (1)$$

where S is a metaphor, P is a replaceable word, M stands for the matching properties of objects denoted by S and P, A is a property of the object S, T is a property of the object P.

It is not difficult to prove that the latter formula is formally contradictory.

The question is as follows: what role does the contradiction play in the metaphor? Part of the answer to this question can be found in the work of P. Ricoeur [6]. He believes that the contradiction in metaphors creates tension between terms, which is the essence of metaphorical meaning.

Hence, in order to increase this «tension» and, accordingly, the aesthetic attractiveness of a metaphor, the difference in values between the metaphor word and the replaceable word must be as large as possible; not just different gradations of the values for one property (for example, like «architectural element» and «tool « in the above metaphor), but the values should be close to the level of antonyms (a small needle and a huge spire). This is one of the main features of metaphors, in which a «strong» contradiction is a necessary component.

From the point of view of logical analysis, this situation occurs not only in metaphors. For example, in reasoning by analogy, properties of one object are matched with properties of another object based on coincidence of certain properties. Similar identification occurs in some models of case-based reasoning [8]. With this in mind, it makes sense to generalize the paradox arising in metaphors to the numerous cases of matching for various objects by means of replacing their names. To do this, let us consider the paradox of identification. Assume there exist an initial object O and its analogue A, and these objects have a common set of properties PC. The object O is also known to have properties PO, and object A has some incompatible properties PA, which can be expressed by the formula $PA \supset \neg PO$. Then the logical model of identification can be expressed using the formula

$$A \wedge (A \supset PC) \wedge (O \supset PC) \wedge (A \supset PA) \wedge (O \supset PO) \wedge (PA \supset \neg PO) \wedge (A \supset O),$$

in which the subformula $A \supset O$ denotes the procedure for replacing the original object with an analogue, and the subformula A at the beginning is an assertion of the trueness of the analogue. It is easy to verify that this formula, as well as its similar formula (1), is contradictory.

The paradox of identification does not refute frequently encountered and very useful reasoning by analogy or case-based reasoning. This paradox is valid only in cases where the original object and its analogue are identified and investigation reveals incompatibility of some their properties.

3 Presupposition

It is easy to find the term "presupposition" (the term "assumption" is preferably used in papers in English) in publications on logic and philosophy [9 – 15], linguistics [16], cognitive science [17], neuro-linguistic programming (NLP) [18, 19], etc. This concept has several differing definitions. We will hold this one: Presupposition is an assertion stipulated (or considered to be true) when analyzing a major assertion or question, while negation or falsity of the major statement does not influence trueness (or falsity) of the presupposition.

For example, the proposition "John has returned back in his family" presupposes that he had gone from the family one day. Evidently, the phrase inverse to the major statement, i.e., "John has not returned to his family") does not change the trueness of this presupposition. Conversely, the sentence "Peter had enough money to buy a smartphone" cannot be correctly used as a presupposition for the proposition "Peter bought a smartphone in a store" since the negation of the major statement, namely "Peter did not buy a smartphone", can be caused in particular by the reason that Peter did not have enough money to pay at that time. Presuppositions are often included in the major statement explicitly. For instance, the phrase "Richard did not know that wolves were found in this forest" clearly presupposes that "There are wolves in this forest."

Hidden presuppositions often cause subconscious perception of some assertions. Sometimes, this is used to manipulate attitudes of people, i.e., to manipulate consciousness [18]. The same can be applied for advertising and in disputes for asking "tricky" questions implicitly presupposing a misdeed of the adversary. Such questions can look like "Do you continue to beat your father?" or "Are you going to return the stolen goods?".

The concept of presupposition was also studied in detail by E. V. Popov [20] during his research in artificial intelligence (AI). His study of communication

with a computer in natural languages displayed that omitting presuppositions during automatic translation can lead to distortions in meaning of texts. In [21], D. A. Pospelov showed how important is to consider presuppositions in models of inference. Modern publications on AI rarely deal with presuppositions. For instance, this concept is absent in fundamental AI monographs like [22, 23]. In accordance with [11], we will further name a major statement as assertion.

It is easy to find the term "presupposition" (the term "assumption" is preferably used in papers in English) in publications on logic and philosophy [9 – 15], linguistics [16], cognitive science [17], neuro-linguistic programming (NLP) [18, 19], etc. This concept has several differing definitions. We will hold this one: Presupposition is an assertion stipulated (or considered to be true) when analyzing a major assertion or question, while negation or falsity of the major statement does not influence trueness (or falsity) of the presupposition.

For example, the proposition "John has returned back in his family" presupposes that he had gone from the family one day. Evidently, the phrase inverse to the major statement, i.e., "John has not returned to his family") does not change the trueness of this presupposition. Conversely, the sentence "Peter had enough money to buy a smartphone" cannot be correctly used as a presupposition for the proposition "Peter bought a smartphone in a store" since the negation of the major statement, namely "Peter did not buy a smartphone", can be caused in particular by the reason that Peter did not have enough money to pay at that time. Presuppositions are often included in the major statement explicitly. For instance, the phrase "Richard did not know that wolves were found in this forest" clearly presupposes that "There are wolves in this forest."

Hidden presuppositions often cause subconscious perception of some assertions. Sometimes, this is used to manipulate attitudes of people, i.e., to manipulate consciousness [18]. The same can be applied for advertising and in disputes for asking "tricky" questions implicitly presupposing a misdeed of the adversary. Such questions can look like "Do you continue to beat your father?" or "Are you going to return the stolen goods?".

The concept of presupposition was also studied in detail by E. V. Popov [20] during his research in artificial intelligence (AI). His study of communication with a computer in natural languages displayed that omitting presuppositions during automatic translation can lead to distortions in meaning of texts. In [21], D. A. Pospelov showed how important is to consider presuppositions in models of inference. Modern publications on AI rarely deal with presuppositions. For instance, this concept is absent in fundamental AI monographs like [22, 23]. In accordance with [11], we will further name a major statement as assertion.

4 Logical Analysis of Presuppositions

Connection problems between explicit and implicit information stimulated logicians in the Middle Ages already [21]. Conversely, linguists consider G. Frege one of the first researchers who drew scientific attention to hidden statements in logical analysis. Particularly, he analyzed distinctions between the assertion in a statement and presupposition(s) for this assertion [24]. He understood presuppositions fairly simply yet, namely only as statements about existence of a referenced entity. For instance, he regarded existence of a person Mozart as an only presupposition for the phrase "Mozart died in poverty".

P. Strawson [9] and B. van Fraassen [10] logically analyzed presuppositions in detail. Strawson proposed the following definition for presuppositions: a sentence P is a presupposition of S, if trueness of P is a necessary condition for S to be true (i.e., S can be either true or false). If P is false, S has no value.

Van Fraassen studied relationships between presupposition and implication. One his proposed definitions stated as follows:

P is a presupposition of S, if and only if:

(a) if S is true, then P is true,

(b) if (non-S) is true, then P is true.

In propositional calculus, this definition corresponds to the formula:

$$\text{P is a presupposition of S if } (S \supset P) \text{ and } (\neg S \supset P).$$

Some studies define presupposition by using the notion of "corollary" rather than implication: "A statement P is a presupposition of S, if it is a corollary from both S and from the negation of S." However, such definitions can cause problems. For instance in [11], the author noted that the formula $(S \supset P) \wedge (\neg S \supset P)$ is equipotent to P, that is such interpretation yields fictitiousness of S. Moreover, many examples of presuppositions show that it usually is a precondition for an assertion, and the opposite interpretation is wrong. Events used in an assertion are mostly a prolongation of the events comprising a presupposition, so the former events may not be considered as preconditions/antecedents.

Sometimes, logical analysis for obtaining a presupposition P for a given statement S reminds derivation a corollary. For example, the reasoning: "The fact John used to beat his father (P) can be derived from the fact that John continues to beat his father (S)" looks likely. However, the more thorough analysis shows that here we have restoring of a former event rather than deducing P from S.

The above-stated leads us to logically define the presupposition unlike the mentioned authors do. Suppose we have the assertion (S) "Anthony was late for school." Evidently, the statement (P) "Anthony was going to school" is a presupposition of S. The latter sentence is also true if Anthony was not late. If we

suppose P to be false, S has no sense at all. We cannot consider it false since its negation ("Anthony was not late for school") is actually false too.

The formal approach results in a paradox when presupposition is treated as a precondition. As the formula $(P \supset S) \wedge (P \supset \neg S)$ is equivalent to $\neg P$, the identical falsity of the presupposition can be stated, although it is assumed to be true according to the meaning of the statements. At the same time, no paradox results from informal analysis of all examples of presuppositions. To study this controversy, we need to closer investigate examples of presuppositions. For the assertion "Anthony was late for school", the fact "Anthony was going to school" is obviously a preceding event for Anthony's being late (or not late).

Within classical logic, we have to add a new factor into reasoning in order to explain presupposition as a precondition. Evidently when Anthony went to school, he could have different reasons to be late (oversleeping, meeting with friends and talking with them, helping an old woman in crossing a road, and so on). Conversely, if no such interfering factors occur, Anthony would not be late. So a presupposition can serve a correct precondition of a major assertion, if we introduce one or more new factors (attributes, variables) into reasoning. In our case, this can be a logical variable R clarifying existence or not existence of reasons for Anthony to be late for school. Obviously, such a factor is necessary to substantiate some strange features" of presupposition.

To get rid of the paradox, we formulate the following hypothesis. Let an assertion S and its presupposition P be given, and we add a new variable R called the relay of an assertion. Within propositional calculus, we obtain:

Hypothesis. If P is a presupposition of a sentence S, then there exists and can be found a logical variable R such that the expression $P \wedge R$ is a prerequisite of the sentence S, and the expression $P \wedge \neg R$ is a prerequisite of the sentence $\neg S$.

Then the argument containing the presupposition and the relay of the assertion can be written by the formula $((P \wedge R) \supset S) \wedge ((P \wedge \neg R) \supset \neg S)$.

We can find that this formula contains no dummy variables, hence, all variables (namely, the assertion S, the presupposition P and the relay R) are not fictitious. Moreover, P can become true or false in this formula, so there is no paradox in such reasoning.

This hypothesis can be favored by the fact that the trueness of the presupposition is preserved for any value of the trueness of the assertion. Hence, it seems quite possible to surmise that the assertion changes its values of trueness depending on some other factor(s). In addition, analysis of numerous examples of presuppositions shows that such a factor (i.e., the relay of an assertion) can always be found.

For instance, the sentence "Alex did not pass the contest to an institute" can have a presupposition "Alex tried to enter the Institute." The negation of

the original sentence does not influence the trueness of this presupposition. To substantiate reasons why Alex passed (or did not pass) this contest, we will need at least one more attribute (for instance, the level of Alex's capabilities, term of his practicing, etc.).

On the other hand, this technique does not explain another "peculiarity" of presupposition. Namely, if we false or deny a presupposition, the sense of both assertion and its negation is lost. For example, the sentence "Jones has hitherto been sick" can serve a presupposition for the assertion "Jones has recovered." If we deny the presupposition, neither assertion nor its negation makes sense since Jones did not recover. Other examples of presupposition display the same feature.

Explanation of this phenomenon is assumed to be beyond the binary logic, i.e., a non-classical logic should be used as an analysis tool [11, 15]. However, it is possible to solve this problem within the framework of classical logic by using a rule-based knowledge representation system.

Consider what happens when denying a presupposition. The presupposition is implicitly present in the values of both assertion and its negation (both "somebody was late to school" and "somebody was not late to school" imply that "someone was going to school"). Thus we can explain why negating a presupposition leads to inconsistency of both the assertion and its negation to the meaning of the modified presupposition: the meaning of the presupposition has changed, and the meanings of the assertion and its negation still imply the former meaning of the presupposition. At the same time, we can assume that when changing a presupposition, the assertion shall also be changed in order to either imply the modified presupposition or to skip a premise associated with the new assertion and the initial presupposition. Hence, the new assertion has to admit a value that differs from the values of the initial assertion and its negation. Then the previous values of the assertion are not the negations of each other, but simply incompatible situations (in our example, the value "did not come" becomes possible in addition to the values "was late" and "arrived on time").

If an assertion is represented as a logical variable, many researchers consider non-classical logic a necessary tool for modeling presuppositions. In this model, assertion can take three values rather than two. To remain in the classical framework, it is sufficient to assume that the variables P, R, and S, which respectively stand for presupposition, relay of assertion and assertion itself, are not logical variables, but unary predicates that have 2, 2, and 3 possible values, correspondingly. To model presupposition within this model, we can use a rule-based knowledge base that can be represented by means of predicate calculus.

Let ranges of predicates values be given as: $P = \{0, 1\}$, $R = \{0, 1\}$, $S = \{p, q, r\}$. Then the logical model of presupposition will look like this:

If P is a presupposition of S, the following rules are taking place.

(1) if P(1) and R(1), then S(a);

(2) if P(1) and R(0), then S(b);

(3) if P(0), then S(c).

Rule (3) models the "inanity" of the assertion in case its presupposition is false. Hence, if Jones was not sick that corresponds to P(0), he is neither recovered (¬S(a)), nor continues to be sick (¬S(b)) now. For P(0) situations are possible in which Jones continues to be healthy (S(c)) or he fell ill all of a sudden. We can describe the latter situation, if we first add another possible value c1 to the values of the predicate S. Sure, we do not wish Jones this option.

The above-proposed technique allows analyzing most examples of presuppositions.

5 Relation of Presupposition to Contradiction Anomalies in KBs

Many papers ([25 – 27, etc.]) deal with anomalies in KBs, which formalize some rules of reasoning. Research on KBs' verification [28] yielded some first methods to recognize and eliminate KBs' anomalies, which can relate to integrity violations (wrong definitions of types and values of attributes) or to consistency violations (mistaken rules themselves). Consider one anomaly of the second category, namely the contradiction anomaly.

Usually, a rule rp looks like $B1 \wedge B2 \wedge \ldots \wedge Bn \rightarrow A$. The part to the left of the arrow is called the antecedent of the rule, and the right part is the consequent of this rule.

Contradiction anomaly. Let the following two rules be given:

r1: $B1 \wedge B2 \wedge \ldots \wedge Bn \rightarrow D$;

r2: $C1 \wedge C2 \wedge \ldots \wedge Cn \rightarrow F$.

Besides, $Ci \subseteq Bi$ for each i (i = 1, 2,…, n) and $D \cap F = \varnothing$. In such a case, the listed rules contain an anomaly of contradiction.

Rules with contradicting consequents and coinciding antecedents describe a particular example of contradictory rules. For instance, the KB of a robot can include following two rules:

rp: $B1 \wedge B2 \rightarrow D$;

rq: $B1 \wedge B2 \rightarrow F$.

Here B1 corresponds to the statement "there is an obstacle ahead", B2 states that "the target is behind the obstacle", D advises to bypass the obstacle on the right, and F wants the robot to pass the obstacle on the left.

Both rules are executable and contradict each other nevertheless. If we

consider them contradictory ones, some revising of the KB becomes necessary, in particular, deletion of one of the rules. Conversely, such a removal can result in poor consequences in some cases, when this anomaly does not manifest an error in a KB. Some rules that look conflicting in the course of reasoning can become absolutely correct after considering some data, which had been possibly missed. In this respect, the contradiction anomaly corresponds to the presupposition.

Consider the above-introduced example of two contradicting rules:

rp: if an obstacle is ahead, and the goal is behind the obstacle, go around the obstacle on the right;

rq: if an obstacle is ahead, and the goal is behind the obstacle, go around the obstacle on the left.

As we said this case looks more like a lacking presupposition than an anomaly. So, no change of the KB is needed, rather, we have to define or find an assertion relay (see Definitions 1 and 2), i.e., an additional attribute that describes different obstacles located on the right and left of the main obstacle, namely, their list and locations.

As we saw, contradictory rules and presuppositions have almost the same definitions. However, the precondition of a presupposition includes one variable, and matching (or nested) preconditions of contradicting rules can comprise several attributes. Sometimes, this allows us to consider contradiction anomalies as a sign for searching an assertion relay rather than declaring a paradox. Then it is necessary to finding some additional variables that solve the paradox.

Now, we propose some techniques to search for a relay of an assertion. Let conflicting rules in a KB be given:

rD: $B1 \wedge B2 \wedge \ldots \wedge Bn \rightarrow D$;

rF: $B1 \wedge B2 \wedge \ldots \wedge Bn \rightarrow F$,

and $D \wedge F = \text{False}$.

Here each disjunct Bi is a couple "attribute – its value(s)". For instance, the condition "If Xi = a or Xi = b, then ..." means that Bi includes the set of values {a, b}. The atoms D and F belong to the same attribute, and the equality $D \wedge F = \text{False}$ means that the sets of their values do not overlap.

When describing a system in terms of "attribute – its value(s)", every rule is defined within a certain relation diagram, which is determined by a set of attributes. The relation diagram of the antecedent of a rule rm is named Ant(rm), Cons(rm) will mean the relation diagram of the consequent for this rule, and Val(Xi, rm) is the value of the attribute Xi in this rule. Suppose we have two contradicting rules rD and rF, and XContr is the set of attributes in their antecedents, so Ant(rD) = Ant(rF) = XContr.

To find an assertion relay for the above rules, we propose the following algorithm.

1) Determine the subset S of the rules rm present in a KB for which Cons(rm) = {Y};

2) Within the set S, determine subsets of rules SD $\subseteq$ S and SF $\subseteq$ S with values of the attribute Y equal to Val(Y, rD) and Val(Y, rF) respectively;

3) For the above-obtained subsets SD and SF, remove rules for which the correlations XContr $\subset$ Ant(rD) and XContr $\subset$ Ant(rF) are not satisfied (strict inclusion means that antecedents of the selected rules contain other attributes besides XContr);

4) From the obtained sets SD and SF, exclude the rules in which the attributes values differ from the corresponding values in the conflicting rules rD and rF;

5) From the sets SD and SF, form the set P of rules pairs (rm, rn) such that rm $\in$ SD, rn $\in$ SF, and Ant(rm) $\cap$ Ant(rn)) \ XContr $\neq \varnothing$ (i.e., their antecedents have other common attributes apart from XContr);

6) For every pair (rm, rn) of P and every attribute Xi of the set (Ant(rm) $\cap$ Ant(rn)) \ XContr, check the correlation Val(Xi, rm) $\cap$ Val(Xi, rn) = $\varnothing$;

7) If the intersection in step 6 is empty for an attribute Xi, this attribute is an assertion relay for conflicting rules rD and rF.

If the introduced algorithm gives no positive result, a similar search can be applied for pairs, triples, etc. of attributes to check whether they can serve assertion relays rather than single attributes.

There is a simple example of seemingly contradictory rules in a KB:

(i) if a bull goes towards a man, and the man shows the bull a red rag, then there is a big risk of causing significant damage to the health of the man;

(ii) if a bull goes towards a man, and the man shows the bull a red rag, then there is no risk of causing significant damage to the health of the man.

Consider how this algorithm works. Let X denote the attribute "direction of the bull" with values "to the subject", "from the subject"; Y is the attribute "red rag in the hands of the subject" with values "true" and "false"; and Z means the risk of causing significant damage to the health of the subject with values "big" and "small".

In accordance with the algorithm, we form sets of rules SD and SF, which have Z as the consequent, while its value is high risk in SD and small risk I SF (Steps 1 and 2).

In the sets SD and SF, we retain only rules whose antecedents contain other attributes in addition to X and Y (Step 3).

In the selected rules, we retain only those for which the value of the attribute X is "to the subject", and the value of the attribute Y is "true" (Step 4).

From the sets of rules SD and SF, we form a set P of pairs (rm, rn) of rules with different values of the consequent (attribute Z); these rules can contain the

same attributes Wj in their antecedents and those attributes differ from X and Y (among such attributes, an assertion relay may be found) (Step 5).

In each pair of the set P, we compare the values of the attributes from the set Wj; if for an attribute its values in different rules are incompatible, this attribute is an assertion relay (Steps 6 and 7).

For example, such an attribute may be the location of the subject (the subject, in particular, may be standing in a clear field, or be in the cab of an armored personnel carrier). Another possible option is the attribute "subject's profession"; if he is a bullfighter, the value of the attribute Z is "small", otherwise it is "big".

As we can see from the description of the algorithm, its computational complexity is evidently polynomial (not higher than the second degree of the total number of rules in the knowledge base). It becomes more complex, if the solution of the problem requires for the search for intermediate conclusions (for example, "if the bull goes towards the subject and the subject shows the bull a red rag, then the bull is expected to attack the subject"). This algorithm is under development.

6 Conclusion

For metaphors, we developed a model applicable in propositional calculus and beyond it. This model allows keeping the difference in values between the metaphor word and the replaceable word as large as possible up to the level of antonyms, which is good to increase the aesthetic attractiveness of a metaphor.

For presuppositions, we analyze their connection with contradiction anomalies in knowledge bases. To explain presupposition as a precondition within the framework of classical logic, it is suggested to supply reasoning with an assertion relay that is new factor(s) formalizing presence or absence of reasons for trueness or falsity of an assertion. To simulate the “inanity” of an assertion when its presupposition is denied, it is proposed to use the model of predicate calculus instead of propositional calculus in order to define the assertion as a logical variable with more than two values. For contradicting rules, we introduced an algorithm to determine possible assertion relays.

Acknowledgement.

The authors would like to thank the Russian Foundation for Basic Researches (grants 16-29-04424, 16-29-12901, 18-07-00132, 18-01-00076, 18-29-03022) for partial funding of this research.

References

1. Conradie W., Goranko V. Logic and Discrete Mathematics: a Concise Introduction. Wiley, 2015. 426 p.
2. Simovici D. A., Djeraba Ch. Mathematical Tools for Data Mining: Set Theory, Partial Orders, Combinatorics. Springer, 2nd Edition, 2014. 834 p.
3. Kulik B. A. Logic of Natural Reasoning. St. Petersburg: Nevsky dialect, 2001. 128 p. (in Russian).
4. Laguta O. N. Metaphorology: Theoretical Aspects. Part 1. Novosibirsk: Novosibirsk state univ., 2003. 114 p. (in Russian).
5. Arutyunova N. D. Metaphor and Discourse. In: Theory of Metaphor. Moscow: Progress, 1990. Pp. 5–32 (in Russian).
6. Ricoeur P. La métaphore vive. Paris: Éditions du Seuil, 1975. 415 p.
7. Leonenko L. L. Metaphor and Analogy // J. of the Ural State University. Series 3: Social Sciences, No 42, 2006. Pp. 23–34 (in Russian}.
8. Intelligent Control Systems: Theory and Applications / Eds. M.M. Gupta, N.K. Sinha. – New York: IEEE Press. 1996. 820 p.
9. Strawson P. Introduction to Logical Theory. London, 1952. 288 p.
10. Fraassen B. van. Presupposition, Implication and Self-reference // J. of Philosophy, vol. 65, No 5, 1968. Pp. 136–152.
11. Beaver D. Presupposition and Assertion in Dynamic Semantics. Stanford: CSLI Publications. 2001. 305 p.
12. Rooij, van, R. Strengthening conditional presuppositions. J. of Semantics, vol. 24, 2007. Pp. 289–304.
13. Rothschild D. Presupposition Projection and Logical Equivalence // Philosophical Perspectives, vol. 2, Issue 1, December 2008. Pp. 473–497.
14. Gutschera K. D. Logical implications and presuppositions in English complement constructions. Doctoral Dissertation. February 2014. 99 p.
15. Chernoskutov Yu. Yu. Context and Logical Theories of Presupposition. Saint Petersburg, 2005. 238 p. // ojs.philosophy.spbu.ru/index.php/lphs/article/download/135/136 (in Russian).
16. Karttunen L., Peters S. Requiem for Presupposition // Proceedings of the Third Annual Meeting of Berkeley Linguistic Society. Berkeley, 1977. Pp. 360–371.
17. Lakoff G. Women, Fire, and Dangerous Things. Chicago, IL: The University of Chicago Press, 1987. 614 p.
18. Bandler R., Grinder J. The Structure of Magic I: A Book About Language and Therapy. Palo Alto, CA: Science & Behavior Books, 1975. 225 p.
19. Tosey P., Mathison J. The Presuppositions of NLP. In: Neuro-Linguistic Programming. Palgrave Macmillan, London, 2009. Pp. 97–110.
20. Popov E. V. Communication with a Computer in a Natural Language. Moscow: Nauka, 1982. 360 p. (in Russian).
21. Pospelov D. A. Modeling of Reasoning. Experience in Analysis of Mental Acts. Moscow: Radio i Svyaz', 1989. 184 p. (in Russian).

22. Russel S., Norvig P. Artificial Intelligence: A Modern Approach. 3rd ed. Prentice Hall, 2009. 1152 p.
23. Thayse A., Gribomont P., Hulin G. et al. Approche logique de l'intelligence artificielle, vol. 1. De la logique classique a la programmation logique. Paris, 1988. 402 p.
24. Frege G. Sinn und Bedeutung. In: Frege G. Funktion, Begriff, Bedeutung. Fünf logische Studien, Vandenhoeck & Ruprecht, Göttingen, 1962. Pp. 38–63.
25. Preece A. D. Validation of Knowledge-Based Systems: The State-of-the-Art in North America // J. of Communication and Cognition – Artificial Intelligence, vol. 1, No 4, 1994. Pp. 381–413.
26. Felfernig A., Friedrich G., Jannach D., Stumptner M.: Consistency-based Diagnosis of Configuration Knowledge Bases // AI Journal, vol. 152, No 2, 2004. Pp. 213 – 334.
27. Baumeister J., Seipel D. Anomalies in Ontologies with Rules // Web Semantics: Science, Services and Agents on the World Wide Web, vol. 8, No 1. 2010. Pp. 55–68.
28. Nguyen T. A., Perkins W. A., Laffey T. J., Pecora D. Knowledge Base Verification // AI Magazine, vol. 8, No 2, 1987. Pp. 69–75.

An analysis of plane task text ellipticity and the possibility of ellipses reconstructing based on cognitive modeling geometric objects and actions

Xenia Naidenova[1], Sergei Kurbatov[2] and Vjacheslav Ganapol'skii[3]

[1] Military Medical Academy, Saint Petersburg, Russian Federation
E-mail: ksennaidd@gmail.com
[2] Research Center of Electronic Computer Engineering, Moscow, Russian Federation
E-mail: curbatow.serg@yandex.ru
[3] Military Medical Academy, Saint Petersburg, Russian Federation
E-mail: Ganvp@mail.ru
lncs@springer.com

Abstract. The article describes the processing of ellipses in an automated system of solving planimetric tasks according to their description in natural language. An approach is proposed to processing ellipses basing on cognitive semantics. The resolution of ellipses is based on using syntactic structures and semantics of geometry in parallel. The types of ellipses most frequently encountered in geometric tasks are revealed. A new approach to recognizing and resolving ellipses in the framework of cognitive semantics is offered.

Keywords: ellipsis resolution, cognitive semantics, planimetric task, text understanding.

2 Introduction

The ambiguity of natural language caused by homonymy has long been studied by computer linguistics. However, the ambiguity associated with the omission of a thinkable language unit (ellipsis) in text has been actively analyzed in natural language processing relatively recently [1], [2]. Although in theoretical linguistics ellipticity got enough coverage [3], [4], restoration of ellipses in systems of syntactic text analysis is clearly developed not enough. Firstly, this is largely due to the fact that eliminating ellipticity is subordinate to actual syntactic analysis and, secondly, this is caused by complexity of resolving ellipses.

The complexity is explained by the necessity to consider a number of contexts: current sentence, adjacent sentences, already established syntactic relations and, finally, semantics of the text. This work is divided into two parts. In the first part, it is described how to handle ellipticity in a specific holistic system of solving plane geometry tasks described in natural language. This system has been implemented in the framework of the INTEGRO project (INTEGRating Ontology) [5]. The second part proposes a new approach to the processing of ellipses based on cognitive semantics.

2 Resolving ellipses in the texts of geometrical tasks

2.1 Syntactical analysis

The architecture and principles of functioning of the system for solving geometrical problems are described in [6] and its general scheme is illustrated by Fig. 1.

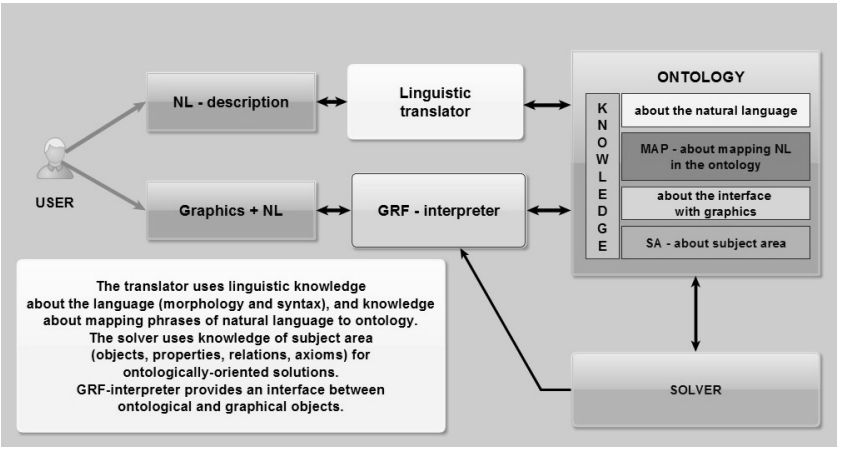


Fig. 1. Scheme of the system for solving geometrical tasks

The system includes the following blocks: linguistic translator, ontology, solver, and graphical module for displaying and explaining the results (drawing NL-explanation of the solution process). The solver receives the ontological structure of the task and forms a chain of basic operations using knowledge of the subject area. In this section, we concentrate on the extension of the system to correctly interpret elliptical (incomplete) sentences.

The language translator creates a syntactic structure and determines that some of its elements violate the language rules. For example, there is no noun for the adjective, the pretext is at the end of the sentence, the number does not have a mandatory measuring unit, and so on. The basic criteria for determining ellipticity are studied by linguists [7, 8]. Based on these criteria recorded in the ontology, the translator identifies the fragments of the syntactic tree that admittedly contain ellipticity. Next, with the use of algorithms described in short below in section 2.2, the identified ellipses are restored. Specifically, in sentence "the radius of the first circle equals 12 cm, and the second 10 cm", the elements "second" and "10" define the ellipticity. As a result, two syntactical structures are formed:

•The radius of the first circle equals 12 cm;

• The radius of the second circle equals 10 cm.

These structures are further processed by the system mechanisms of paraphrasing to obtain an ontological representation of sentence in the formal terms of the subject area [6]. The concept "paraphrasing" has been proposed by the well-known Russian linguist Apresyan in [9]. In our system, we use an adaptive variant of this concept. The conception of paraphrasing assumes that any class of sentences corresponding to one and only one sense can be reduced to the simplest or canonical phrase composed only of the lexemes expressing most clearly the basic concepts of sentences. Thus, paraphrasing is based on the following proposition in [9]: "One of the fundamental properties of human languages consists in the fact that if there are several synonyms, in the broad sense, to express some concept, then only one of them turns out to be privileged, canonical, or prototypical for expressing this concept". In particular, such canonical concepts in plane geometry are, for example, the point, the line, the plane and to belong, to lie between, and to be congruent. Thus the rules of paraphrasing provide only one canonical form for a group of sentences having the same sense. For example, sentences "a point located on the straight line", "the straight line passing through a point", "a point belonging to the line", "a point lying on the line segment" etc. are reduced to the following canonical phrase "point belongs to straight line". This canonical phrase is mapped to its ontological representation in the form of the following triplet "point lies line". It should be stressed that the members of the triplet (objects and relations between them) are not dependent on a language. Therefore the corresponding rule of paraphrasing contains, in its left part, the objects and relations depending on language, but, in its right part, the formal objects and relations invariant in different languages.

The rules of paraphrasing are divided into two classes; the first one consists of rules in which both parts are some generalized syntactic structures; the second one consists of rules having canonical descriptions in their left parts and semantic descriptions in their right parts. The second class of rules can be

used for transforming ontological structures into corresponding natural language texts. It is reasonable to apply the rules of the first class to equivalent synonymic transformations of synthesized structures to retrieve texts in the most appropriate manner in a considered application domain.

2.2 Algorithm for resolving ellipticity

The algorithm for treating ellipses is based on the ontology knowledge reflecting the semantic hierarchy of word forms in the syntactic structure and the norms of natural language. To a first approximation the algorithm can be described as follows:

- to segment a syntactic structure into two segments: a complete one without ellipticity and the other one containing ellipticity (generally, it is a set of noun groups (NG));
- in the elliptical segment, to reveal the elements that are supposed to be used for resolving ellipticity;
- in the full syntactic structure, to reveal the candidates to be replaced by the elements found in the previous step;
- to perform the replacement and obtain the complete syntactic structure.

In the example given in section 2.1 "*first*" is replaced by "*second*" and "*12*" by "*10*" because they correspond to the same concepts of ontology. Here we have different objects and the same type of attribute (length). In the sentence "*the perimeter of triangle is 37 cm and the area − 20 cm*" we have the same object and different types of attributes. This seemingly simple algorithm allows to successfully recover not only geometrical ellipses, but several others, described, for example, in [2]: in the sentence "*twenty years of such dance form the age, forty − the history*" "*twenty*" is replaced by "*forty*" and "*age*" is replaced by "*history*".

2.3 Limitations

Of course, many cases of ellipticity cannot be processed by the algorithm above. Example: "There are seven circles. Radius of one 5 cm, two others − 3 cm, and the others − 10 cm". We have multiple ellipticity in this example. A similar example from [2]: "Anemones discard tentacles, crabs − claws, lizards − tail". In many cases, ambiguity arises at the level of comparison. Two options were analyzed: 1) to move forward with analyzing the situation and eliminating ambiguity at the stage of semantic processing; 2) to complement the ontology by the rules of preferences when choosing a candidate for replacement (substitution). It should be noted that the question of clear ellipticity criteria and methods for restoring

the full structure of sentences has not been fully resolved within the framework of a generally accepted linguistic theory. Resolving ellipses in natural language texts remains one of the most difficult and unsolved tasks in linguistics, despite the abundance of proposed methods based on syntactic-semantic parsing of sentences. Syntax reveals the structure of the ellipsis and the similar part of the sentence without it; semantics deals with word values. However, as the example from [11, page. 62] shows, resolving ellipses is based on the understanding of context (text theme), the sense of words and collocations: "Charles makes love with his wife twice a week. So does John".

2.4 Testing the algorithm

The algorithm performs the ellipses' resolution with the accuracy equal to 100% in simple cases when the noun phrase in a sentence consists of only one word. It is important to note that resolving ellipses is directly connected with the correct functioning the system ontology, since the ontology supports the process of sentence understanding. In more complex cases with the composite noun phrases or incomplete ontology, the accuracy of the algorithm declines to 70 %. In any case, difficult texts of some planimetric tasks require the special analysis and the solution ad hoc.

Currently, several hundred of simple ellipses and several tens of complex ones have been tested.

In general, it should be anticipated that the vast majority of sentences contains several types of ellipses or some number of ellipses of the same type. This fact implies the search for some new approach to reconstructing ellipses covering not only the ontology and linguistic knowledge but also the model of human plausible reasoning and cognitive model of practical geometrical situations. Ellipsis resolution must be based on cognitive semantics.

3 Ellipsis classification in geometrical tasks

To study the typology of ellipses in geometric tasks we used a body of texts containing more than 1000 planimetric tasks. We have revealed the following types of ellipses: ellipses using dash "–" (ellipses with skipped predicate or verb), ellipses without "–" (ellipses with skipped verb, noun, pronoun, or predicate. Consider the structure of these ellipses. We will give only fragments of tasks containing ellipses.

Skipped predicate: *In triangular ABC there are given R and r – radii of circumscribed and inscribed circles. A_1, B_1, C_1 – points of crossing the bisectors of triangle ABC with the circumscribed circle.*

Structural components of these ellipses are Noun Phrase (NP) and Prepositional Phrase (PP). Revealing NP and PP is realized in the system OntoIntegrator [12] in the framework of the project on creating World Digital Mathematical Library – WDML.

Consider this type of ellipses in greater detail:

a) < NP > < – > < Designation(s) (Bases of perpendiculars dropped from B and D on AC – M and N);

b) < Designation(s) > < –> < NP > (O1, O2, O3, O4 − centers of circles; D – arbitrary point of the plane; BD – the side of rectilinear pentagon inscribed in this circle);

c) < NP > < – > < NP > (The points of their intersection lie on the same circle – the circle of nine points; This quadrangle is a diamond; Every parallelogram inscribed in a circle – rectangle; Every diamond inscribed in a circle – square);

d) <NP> < – > <PP> (Center of circle – inside the quadrangle; C – between A and F);

e) < NP > < – > <The property expressed by adjective> (Angle C – right; To find a point on a given line such that the sum of distances from it to two points A, B – minimal).

The resolution of these ellipses can be carried out according to the scheme:

to select NPs; to identify the heads of NPs as geometrical objects; to identify designations; to localize the dash between the designation(s) and the NPs; to check (according to the rules of the ontology) the conformity between the designations and the heads of the NPs; to restore ellipses. In these cases, the dash is replaced by the forms “is” or “are” of the verb “to be”.

The dash in the Russian language is put in a variety of situations. In situation c), the dash is put between the subject and the predicate in the absence of a link between them [13], if both members are expressed as nouns in the form of the same case, for example, “Loneliness in a creative work – a hard thing”, “The next station is Mytishchi”. In geometric problems, situation c) has the nature of a logical definition (geometry – a section of mathematics) or identity, when the subject and the predicate are expressed by the same concept. If the subject and the predicate are not expressed by the same word, then it is necessary to check the predicative relation through logical inference in the ontology.

In view of our consideration of Verb Phrase Ellipsis in the previous section we confine ourselves to one of difficult cases of this ellipsis.

Skipped verb (ellipsis with dash): *In triangle ABC there are taken points M, N and P: M and N − on sides AC and BC, P − on line segment MN.*

In this sentence, we have an incomplete VP: *In triangle ABC there are taken points M, N and P* (presupposition), this VP is prolongated by the follow way:

In triangle ABC there is taken point M on side AC;
In triangle ABC there is taken point N on side BC;
In triangle ABC there is taken point P on line segment MN.

Restoration of this sentence is supported by a thinkable geometric situation, (let us call it a **cognitive model of a geometric situation**). And the restoration goes on sequentially, but with simultaneous creation of different relationships: temporary (earlier, later), referential (the designation refers to an object, the pronoun refers to an object), spatial (in the triangle, on the side), linguistic (links of relationships, objects, properties with certain word forms and expressions), quantitative. So, in our example we have (→ means a reference):

In triangle ABC → triangle → designation = ABC;
Triangle ABC → one → it → it is given → this → in it;
Triangle ABC → side AC → one, side BC → two, side AB → three
In triangle ABC there are taken points M, N, and P;
Point one → designation M → first, point two → designation N → second;
Point three → designation P → third;
In triangle ABC there is taken point M; in triangle ABC there is taken point N; in triangle ABC there is taken point P;

Now we need a model of acting: "to take point in a tringle" and **generating hypotheses** "Where?". In accordance with one of the hypotheses the following cases are:

In triangle ABC there is taken point M (one) on side AC (one);
In triangle ABC there is taken point N (two) on side BC (two);
By analogy:
In triangle ABC there is taken point P (three) on line segment MN.
Line segment → designation MN → it joins points M and N (supported by knowledge about how a segment of a line is generated).

As a result, we can restore the full text of this task: *In triangle ABC there is taken point M on side AC*; *there is taken point N on side BC***,** *and there is taken P on line segment MN.*

The process of binding objects during their construction is supported by cognitive models of objects and operational knowledge. As D. Suleymanov [14] noted, "it is necessary to go not from the text, but from the task". All cognitive models can be explicitly defined based on geometric semantics and they are associated with speech parts and typical collocations with their grammatical categories at the sentence level.

Restoration of the full text requires reasoning by analogy and understanding the meaning of actions with geometric objects. Exactly, similar actions are supposed with similar objects, and therefore the words are skipped. In practice, most skipped words are redundant for understanding the sense of sentences.

People omit words consciously. However, if the missing information is not redundant, understanding texts represents a problem that is resolved by analyzing geometric situations.

The following sentences give the examples of ellipses without dashes.

Skipped verb, ellipsis without dash: *The vertices of parallelogram $A_1B_1C_1D_1$ lie on the sides of parallelogram ABCD: point A1 lies on AB, B_1 on BC, etc.).* (Word "*lies*" after B_1 is skipped)

Skipped noun: *Prove that the value of angle with the vertex inside a circle equals the half-sum of the angular values of two arcs of which one is enclosed between the sides of this corner and the other between the prolongation of sides.* (Words "*of this corner*" after word "*sides*" are skipped).

Skipped pronoun: *In a circle of radius R, two chords AB and AC are drawn. On AB or on its extension, point M is taken. Analogically, on AC or on the extension, point N is taken.* ("*of it*" is skipped after "*the extension*").

Skipped predicate: *Side BC of triangle ABC is equal to a, radii of a circumscribed circle r.*

4 The structure of cognitive models of objects and actions

Cognitive structures correspond to the semantic structures of situations described in the text. They should be aligned with the narrative structures of sentences. A word can have multiple values, but only one sense, at least in mathematical texts. Ellipsis (omitting words, economy of text) is possible because the preceding text determines unambiguously (uniquely) the meaning of each word and situation, and these meanings remain unchanged. In cognitive models of objects, the following relationships are important:

- object can perform some actions;
- object can be subjected to actions of other objects;
- object can have spatial and temporal relationships (earlier, later, already built, already given) with other objects;
- object can be composed of some other objects;
- object can be a part of some other object (objects);
- object has properties, some of which (call them actant ones) are related to the actions that the object commits (intersects – intersecting, lies – lying) or the actions that are committed over it (has been given – given, has been formed – formed, cut of, embedded). Thus, the actant properties of objects are directly displayed in the morphological forms of words describing these properties;
- the relationships between the properties of one geometrical object and the properties of others.

These relationships are in agreement with the universals described by D. Suleymanov [15]. The properties between an object and its parts are realized through implications: if center, then a circle; if radius, then a circle; if circle, then circumscribed about or inscribed in; if inscribed in, then in an object; if bisector, then bisector of an angle; if bisector of angle, then the vertex of angle from which it originates; if bisector, then the angle from which it comes is divided in half; if bisector, then it is the axis of symmetry of angle divided in half by this bisector.

The interaction of cognitive models and the analyzed text should provide the principle of "cognitive expectation" and "determinism of context" [14].

Creation of cognitive models of objects and actions for plane geometry, in the proposed approach, is performed in a step-by-step mode by the use of a given text corpus. Some fragments of cognitive model "Bisector" are shown in Tables 1 and 2. It is also a problem of considerable interest to apply a plausible reasoning for resolution of ellipses, including analogy, generalizations, specialization, use of implications, forming hypotheses and many others.

Table 1. Noun Phrases with "Bisector"

Bisector	Hyperlink to object (to NP)	Hyperlink to object (to PP)
Bisector of	angle	
Bisector of	angle	in (of) triangle
Bisector of	acute angle	in (of) rectangular triangle
Bisector of	inner angle	in (of) triangle
Bisector of	angle	at base of isosceles triangle
Bisector coming from	vertex	of inscribed triangle
Bisector of	angles adjacent to one side	in (of) parallelogram
Bisector of		in (of) triangle
Bisector of	inner angle	in (of) parallelogram
Bisector of	angle	in (of) convex quadrilateral
Bisector of	angle	in (of) rectangle

Table 2. Verb Phrases with "Bisector"

Bisector	Hyperlink to object (to NP)	Hyperlink to object (to PP)
Dividing	To divide	Side of triangle
Perpendicular	To be perpendicular	Median of triangle
Splitting, cutting in	To split, to cut in	Side of parallelogram in segments

Intersecting	To intersect	Bisector of triangle
Intersecting	To intersect	Circle
Restricting	To restrict	Area of quadrangle
Coming across	To come across	Circle in points
Containing	To contain	Points of intersection
Lying on	To lie	Straight line

Within the proposed approach, text analysis becomes cognitive-driven, and the parser plays a subordinate role (Fig. 2). If ellipsis resolution is based on cognitive models, then it is possible to synthesize a text describing a geometric situation and compare this text with the text to be analyzed. The ontology contains theoretical knowledge in the area to solve geometry tasks of various types (computational, for construction, for proof). The ontology takes the burden of solving tasks and visualizing solutions. The Cognitive Analyzer runs incrementally and transmits a converted and meaningful text to the ontology in the form required by it.

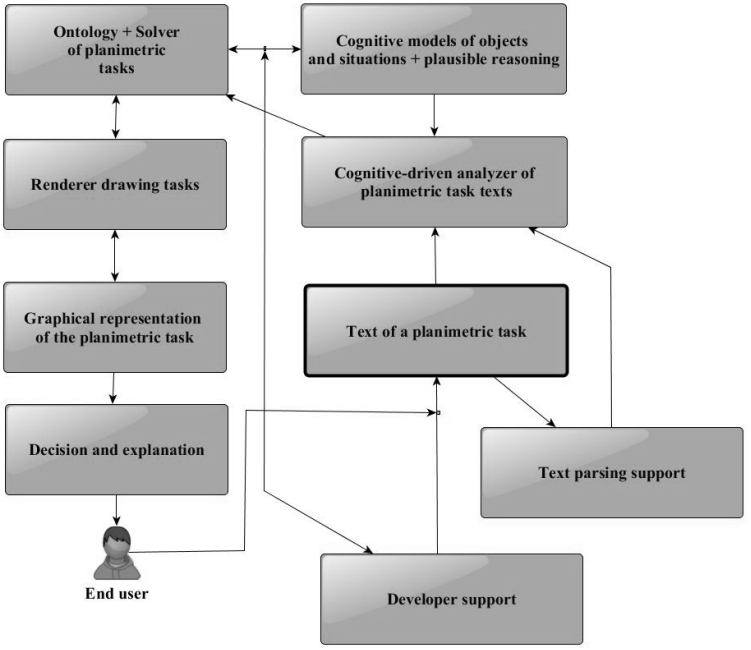


Fig. 2. Scheme of a cognitive-controlled analysis of a text

5 Related works

Verb Phrase Ellipsis is a well-studied topic in theoretical linguistics but has received little attention as a computational problem and a task of human reasoning except the paper [16]. Exhaustive linguistic analysis of ellipses for different languages performed in many sources: for example, [8], [17- 24].

In spite of the fact that a lot of works deal with resolution of ellipses, the significant results are obtained only for some special types of them, namely for the verb ellipses (VE) in the framework of syntactical-semantic analysis.

Detection and resolution of Verb Phrase Ellipsis (VPE) are considered in the articles [25-30] but only for some special cases: resolving elided scopes of modality and ellipses with auxiliary verbs. In [26], the authors have proposed a method of automatic ellipsis resolution without preliminary processing or annotation of texts. This work is carried out within the OntoSem language processing system of the OntoAgent cognitive architecture. OntoAgents carry out deep semantic and pragmatic language analysis, yielding ontologically grounded text meaning representation that populate agent memory and subsequently support agent reasoning [27].

The text with the VE has the following structure consisting of 2 parts standing on the right and left of the "dash" (both parts are in the same sentence). The verb is skipped in the right part, the left part (the antecedent) contains the verb. The right part is complemented by the verb from the left part. Example: *She can **go** to Hawaii but he can't (She can go to Hawaii but he can't go)*.

The resolution of such an ellipsis consists of three stages:

- Recognizing the occurrence of ellipsis, localizing it, and selecting its parts;
- finding the nearest to the left verb in the antecedent;
- resolving ellipsis.

The paper [28] describes a system ViPER (VP Ellipsis Resolver) that detects and resolves VP ellipsis, relying on linguistic principles such as syntactic parallelism and modality correlations. The system ViPER has been incorporated into the OntoSem2 incremental semantic analysis system that provides language analysis capabilities to OntoAgents.

In [27], a novel approach is presented to detecting and resolving VPE by using supervised discriminative machine learning techniques trained on features extracted from an automatically parsed, publicly available dataset. Additionally, this approach uses the Margin-Infused-Relaxed Algorithm for antecedent identification. It is proposed a decomposition of the overall resolution problem into three tasks − target detection (ellipsis detection), antecedent head resolution, and antecedent boundary detection.

The features used for antecedent head resolution and/or boundary determination try to capture aspects of both tasks. The features are roughly grouped by their type. **Labelsfeatures** make use of the parsing labels of the antecedent and target; **Treefeatures** are intended to capture the dependency relations between the antecedent and target; **Distancefeatures** describe distance between them; **Matchfeatures** test whether the context of the antecedent and target are similar; **Semanticfeatures** capture shallow semantic similarity; there are a few **Otherfeatures** which are not categorized.

In [30], a new method is proposed to resolve multiple ellipses in such sentences as:

- Unemployment has reached 27.6% in Azerbaijan, 25.7% in Tadzhikistan, 22.8% in Uzbekistan, 18.8% in Turkmenia, 18% in Armenia and 16.3% in Kirgizia;

In this paper, sentences lack an overt predicate. The authors present two methods for reconstructing elided predicates within the Universal Dependencies (UD) framework. The first method adapts an existing procedure for parsing sentences with elided function words [31], which uses composite labels that can be deterministically turned into dependency graphs in most cases. The second method is a novel procedure that relies on the parser only to identify a gap. Then an unsupervised method is used to reconstruct the elided predicates and reattach the arguments to the reconstructed predicate. The both methods work with very high accuracy (from 81,69 to 90,57 %) and significantly exceed the recently proposed constituent parser by Kummerfeld and [32]. The types of ellipses reconstructed are:

(1) Single predicate gaps:

John **bought** books, and Mary____ flowers.

(2) Contiguous predicate-argument gap (including ACCs):

Eve **gave flowers** to Aland Sue_____ to Paul.

Eve **gave** a CD to Al and____ roses to Sue.

(3) Non-contiguous predicate-argument gap:

Arizona **elected** Goldwater **Senator**, and Pennsylvania_____ Schwelker____.

(4) Verb cluster gap:

I want to try to begin to write a novel and ... Mary _____a play. ...

Mary _____to write a play. ...

Mary _____to begin to write a play. ...

Mary _____to try to begin to write a play.

The core characteristic of resolving ellipses is that there is a clause that lacks a predicate (the gap) but still contains two or more arguments or modifiers of the elided predicate. In most cases, the remnants have a corresponding argument

or modifier in the clause with the overt predicate. The UD frame work aims to provide cross-linguistically consistent dependency annotations that are useful for NLP tasks. The UD defines two types of representation: the basic UD representation which is a strict surface syntax dependency tree and the enhanced UD representation [33] which may be a graph instead of a tree and may contain additional nodes.

See [34] and [35] for a more comprehensive overview of cross-linguistically attested gapping.

The major advantage of this approach is that the dependency tree contains information about the types of arguments and so it should be straightforward to turn dependency trees into enhanced UD graphs. For most dependency trees, one can obtain the enhanced UD graph by splitting the composite relations into its atomic parts and inserting copy nodes at the splitting points.

A crucial step is the third step, determining the highest-scoring alignment. This can be done with the algorithm presented by Needlemann and Wunsch [36] in which one defines a similarity function *sim*(g,f) that returns a similarity score between the arguments g and f. Defining *sim* based on the intuitions that often, parallel arguments are of the same syntactic category, that they are introduced by the same function words (e.g., the same preposition), and that they are closely related in meaning.

Seeker et al. [31] compared three ways of parsing with empty heads: adding a transition that inserts empty nodes, using composite relation labels for nodes that depend on an elided node, and pre-inserting empties before parsing. These papers all focus on recovering nodes for elided function words such as auxiliaries; none of them attempt to recover and resolve the content word elisions of gapping.

6 Conclusion

Processing ellipses is given in a specific system of plane geometry tasks described in natural language. Ellipsis resolution is based on using in parallel the syntax structures of sentences and the geometry semantics. A broader approach to ellipses processing based on cognitive semantics has been proposed. The approach gives a classification of ellipses (across a geometric text corpus) and introduces the concept of a cognitive model of geometry objects and actions. This approach allows to view the structure of automated analysis of geometric texts as a cognitively controlled parsing.

Acknowledgments.

The research was partially supported by Russian Foundation for Basic Research, research project No. 18-07-00098A.

References

1. Malkovsky, M. et al.: Restoring ellipse as a task of automatic text processing. Program Products and Systems 3(107), 32-36 (2014) (in Russian).
2. Kobzareva, T., Epifanov, M., Lakhuti D.: Finding the antecedent of the ellipsis of a fragment containing predicate (automatic analysis of Russian sentence). In: Proceedings of 15th National Conference on Artificial Intelligence with International Participation (CAI-2016),vol. 2, pp. 56-62 (2016) (in Russian).
3. Adametc, P.: Semantic interpretation of "significant zeros" in Russian sentences. Language and verse in Russia. In: Proceedings of papers in honor of Dean C. Worth for his 65 Anniversary, pp. 9-18. East Literature, Russian Academy of Sciences, Moscow (1995) (in Russian).
4. Vardul, I.: To the question of the ellipse phenomenon. In: Proceedings of the conference "Invariant Syntactical Values and Structure of Sentence", pp. 59-70 (1969) (in Russian).
5. Kurbatov, S., Naidenova, X., Khakhalin G.: Integrating intelligent systems of analysis/synthesis of images and text: project outlines INTEGRO. Contours of the INTEGRO project. In: Proceedings of the International Scientific Conference "Open Semantic Technologies for Intelligent Systems" (OSTIS-2011), pp. 213-232 (2011).
6. Kurbatov, S., Vorobyev, A.: Ontological solver of geometry problems in natural language description. In: Proceedings of 15th National Conference on Artificial Intelligence with International Participation (CAI-2016), vol. 1, pp. 56-63 (2016) (in Russian).
7. Larkina, A.: About the lingustic investigations of ellipcises in French languauge. Bulletin of the Chelyabinsk State University 5(143), 7-29 (2009) (in Russian).
8. Minjaylov, V.: The peculiarities of elliptical construction in Russian language. World of Science and Innovation 1(1), 27-31 (2015).
9. Apresyan Y., Boguslavsky, I., Iodmin, L., et al.: Linguistical maintenance of the system ETAP-2. Nauka, Moscow (1989) (in Russian).
10. Lobzin, À., Khakhalin, G., Kurbatov, S., Litvinovich, A.: Integration based on natural language and image ontology in the system Text-To-Picture. In: Proceedings of 8th Scientific-Practical Conference "Integrated Models and Soft Computing in Artificial Intelligence, pp. 296-305 (2015) (in Russian).
11. Eco, U.: The role of the reader. Exploration in the semiotics of texts. Publishing House "ACT": CORPUS, Moscow (2016) (in Russian).
12. Nevzorova, O. and Nevzorov, V.: Terminological annotation of the document in a retrieval context on the basis of technologies of system "OntoIntegrator". International Journal "Information Technologies& Knowledge", 5(2) (2012).
13. Rosental, D.: Reference book on Russian language: orthography and punctuation. Publishing House "ACT", Moscow (2013) (in Russian).
14. Suleymanov, D.: Pragmatic-oriented linguistic models as a basis of systems and technologies for processing natural language. Analytical Review. In: Sosnin P.I. and

Nevsorova O.A. (eds.) "Formal Models and Systems in Computational Linguistics", Chapter 1 (pp. 8 -59). Republic Tatarstan Academy of Sciences, Kazan (2016) (in Russian).
15. Suleymanov, D. and Gatiatullin, A.: A system of semantical universals and their implementation in the form of relational-situational frame. In: Stephanyuk V. and Taysina E. (eds.) "Cognitive-Semiotics Aspects of Modelling in Humanitarian Sphere" (pp. 185 -210). Republic Tatarstan Academy of Sciences, Kazan (2017) (in Russian).
16. Kubota, Y. and Levine, R.: Gapping as hypothetical reasoning. Natural Language and Linguistic Theory 34(1), 107-156 (2016).
17. Kobozeva, I.V.: Linguistical semantics. Publishing House "URSS", Moscow (2000) (in Russian).
18. Mitkov, R.: Outstanding issues in anaphora resolution. In: Gelgukh A (ed.) "Computational Linguistics and Intelligent Text Processing" (pp. 110-125). Springer (2001).
19. Nielsen, L. A.: A corpus-based study of verb phrase ellipsis identification and resolution. Ph.D. thesis. King's College, London (2005).
20. Nielsen, L. A.: Verb phrase ellipsis detection using automatically parsed text. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004). V. 1, pp. 1093-1099 (2004).
21. Bengtson, E., and Roth, D.: Understanding the value of features for coreference resolution In: EMNLP'08 Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.294-303 (2008).
22. Carberry, S.: A pragmatics-based approach to ellipsis resolution. Computational Linguistics 15(2), 75-96 (1989).
23. Carbonell, J. G.: Discourse pragmatics and ellipsis resolution in task-oriented natural language interfaces In: ACL'83 Proceedings of the 21st annual meeting on Association for Computational Linguistics, pp. 64-168 (1983).
24. Ivan, S. and Hankamer, J.: Toward a theory of anaphoric processing. Linguistics and Philosophy 7, 325-345 (1984).
25. McShane, M. & Babkin, P.: Automatic Ellipsis Resolution: Recovering Covert Information from Text. In: Proceedings of the Twenty Ninth AAAI Conference on AI, pp. 572- 578 (2015).
26. McShane, M., Nirenburg, S., Beale, S., Johnson, B.: Resolving Elided Scopes of Modality in OntoAgent. Advances in Cognitive Systems 2, 95-112 (2012).
27. Kenyen-Dean, K., Cheung, J.C.K., Precup, D.: Verb Phrase Resolution Using Discriminative and Margin-Infused Algorithms. In: Proceedings of the 2016 Conf. "Empirical Methods in NLP", pp. 1734-1743 (2016).
28. McShane, M., Babkin, P.: Detection and Resolution of Verb Phrase Ellipsis. LiLT 13(1), 1-36 (2016).
29. Liu, Z., Gonzalez, E., Gillick., D.: Exploring the steps of VPE. In: Proceedings of the Workshop on Conference Resolution Beyond OntoNotes (CORBON 2016), co-located with NAACL, pp. 32-63 (2016).

30. Schuster, S., Nivre, J., and Manning, Ch. D.: Sentences with Gapping: Parsing and Reconstructing Elided Predicates. arXive: 1801.06922v1 [cs.CL] 18 Apr 2018.
31. Seeker, W., Faras, R., Bohnet, B., Schmid, H., and Kuhn, J.: Data-driven dependency parsing with empty heads. In: Proceedings of COLING, pp. 1081-1090 (2012).
32. Kummerfeld, J. K. and Klein, D.: Parsing with traces: An O(n4) algorithm and a structural representation. Transactions of the Association for Computational Linguistics 5, 1-36 (2016).
33. Schuster, S. and Manning, C. D.: Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 2371-2378 (2016).
34. Schuster, S., Lamm, M., and Manning, C.D.: Gapping constructions in Universal Dependencies v2. In: Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pp. 123-132 (2017).
35. Johnson, M.: A simple pattern-matching algorithm for recovering empty nodes and their antecedents. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), 136-143 (2002).
36. Needleman, S. B. and Wunsch, C. D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3), 443-453 (1970).

Named Entity Recognition in Tatar: Corpus-Based Algorithm

Olga Nevzorova[1][0000-0001-8116-9446], Damir Mukhamedshin[1][0000-0003-0078-9198] and Alfiya Galieva[1]

[1] The Tatarstan Academy of Sciences, Kazan, Russia
{onevzoro,damirmuh,amgalieva}@gmail.com

Abstract. Named entities recognition is one of the urgent tasks in the researches of language using electronic language corpuses. This article discusses the main methods for solving this problem, including algorithms based on various machine learning models, regular expressions and dictionaries. Also in the article, the authors proposed their own algorithm, which allows named entities recognition on the basis of search queries using direct and reverse search. The results of the algorithm, presented in the article, suggest what additional functions are necessary to achieve the best results. The proposed algorithm is used in the "Tugan Tel" corpus management system and can be used both with the electronic corpus of the Tatar language and with corpuses of other languages.

Keywords: Named entity recognition, NER, Corpus management system, Text mining.

1 Introduction

Electronic language corpuses are the basis for extensive research related to language research. Corpus management systems help solve a number of linguistic problems, such as direct search of word forms, lemmas, reverse search by morphological properties, selection of contexts, n-grams for various search queries. These simple queries are supported by most corpus management systems.

One of the difficult tasks of searching in corpus data is named entities recognition. This problem is solved by dozens of researchers, often getting good results. Most existing solutions, some of which are described in Section 2 of this article, work with English, Spanish, Dutch, German using various NLP methods,

regular expressions, dictionaries, etc. as the basis. In Section 4 of this article, the authors considered one of the possible algorithms for named entities recognition, which can be used both with the electronic corpus of the Tatar language and with electronic corpuses of other languages. This algorithm is implemented in one of the modules of the "Tugan Tel" corpus management system. The authors also conducted a series of experiments, the results of which are shown in Section 4.2 of this article.

2 "Tugan Tel" Corpus Management System

The Tatar corpus management system (www.corpus.antat.ru) is developed at Institute of Applied Semiotics of the Tatarstan Academy of Sciences. The main functions of the corpus management system are searching for lexical units, making morphological and lexical searches, searching for syntactic units, n-gram searching based on grammar and others. The core of the system is the semantic model of data representation. The search is performed using common open source tools. We use MariaDB database management system and Redis data store [1]. Our purpose is to design the corpus management system for supporting electronic corpora of Turkic languages. This line of research is developing very rapidly.

Among well-known electronic corpora projects for Turkic languages are the corpora of Turkish and Uyghur [2], Bashkir, Khakass, Kazakh (http://til.gov.kz), and Tuvan languages. "Tugan Tel" Tatar national corpus is a linguistic resource of modern literary Tatar. It comprises more than 100 million word forms, at the rate of November 2016. The corpus contains texts of various genres: fiction, media texts, official documents, textbooks, scientific papers etc. Each of the documents has a meta description [3]: author, title, publishing details, date of creation, genre etc. Texts included in the corpus are provided with morphological markup, i.e. information about part of speech and grammatical properties of the word form [4]. The morphological markup is carried out automatically on the basis of the module of two-tier morphological analysis of the Tatar language with the help of PC-KIMMO software tool.

3 Related Works

3.1 LingPipe

One of the related works is LingPipe [5], which is a collection of Java libraries developed by Alias-I. LingPipe allows to classify named entities in English:

person, organization, place. It supports the use of other language packages for classification. LingPipe also supports additional features such as orthographic correction and English text classification. This software is distributed free of charge for research purposes.

3.2 Annie

Another similar work is Annie [6]. This is a named entity extraction module embedded into the GATE framework. Annie is open source and is developed under the GNU license developed at Sheffield University. Annie implements various functions necessary for extracting named entities: tokenizer, sentence separator, POS tagging, resolution with a link, place name directories, etc.

3.3 Afner

Afner [7] is an open source NERC tool licensed under the GNU license, developed in C++ at Macquarie University. It is used as part of a question and answer service that focuses on maximizing responsiveness to user questions. At the same time Afner can be used separately from the service. Afner uses lists, regular expressions, and supervised learning models. It allows one to extract names of persons, organizations, locations, monetary values and dates from English texts.

3.4 Knowledge-based systems

Knowledge-based NER systems use lexical resources and domain-related knowledge without requiring training with annotated data. Such systems show good results when the lexical resources are complete, whereas they do not work, for example, with the examples from drug_n class in the DrugNER [8] data set, since they are not defined in the DrugBank dictionaries. Despite their high accuracy, these systems show low recall due to specific rules of the language and domain and incomplete dictionaries. Another disadvantage of knowledge-based NER systems is the need for experts to participate in the development and maintenance of knowledge resources.

3.5 Unsupervised and bootstrapped systems

Early systems did not require significant data for training. Collins and Singer (1999) [9] used only labeled seeds and 7 functions for classifying and extracting named entities: orthography (for example, capitalization), entity context, words that occurred in named entities, etc. To improve the recall of NER sys-

tems, Etzioni et al. (2005) [10] proposed an unsupervised system using 8 generic pattern extractors for open web texts, for example, *NP is <class1>*, *NP1 such as NPList2*. In 2006, Nadeau et al. suggested using an unsupervised system to create a directory of named entities and resolve the ambiguity of named entities basing on the work of Etzioni et al. (2005) [10] and Collins and Springer (1999) [9]. This system combined the extracted list of named entities with generally accessible directory of named entities and achieved F-scores of 88%, 61% and 59% on MUC-7 [11] for named entities of classes of locations, persons and organizations, respectively.

Zhang and Elhadad (2013) [12] in an unsupervised NER system for biological and medical data used surface syntactic knowledge base and inverse document frequency (IDF). This system reached 53.8% and 69.5%, respectively. Their model uses seeds to find text with possible content of named entities, identifies phrases with nouns and filters phrases with a low IDF value. The filtered list is submitted to the classifier for predicting the tags of named entities.

3.6 Feature-engineered supervised systems

Supervised machine learning models learn to make predictions by training on example inputs and their expected outputs, and can be used to replace humanly established rules. Hidden Markov Models (HMM), Support Vector Machines (SVM), Conditional Random Fields (CRF), and decision trees were common machine learning systems for NER.

The results of research using various machine learning models from various authors are presented in Table 1.

Table 1. Various machine learning models results.

Author(s)	Machine learning model	Additions	Results
Zhou and Su (2002) [13]	HMM	Included 11 orthographic features, a list of trigger words for named entities, and a list of words from various gazetteers.	F-scores of 96.6% and 94.1% on MUC-6 and MUC-7 data, respectively.
Malouf (2002) [14]	HMM and Maximum Entropy (ME)	Included capitalization; considered whether the word went first in the sentence, whether the word had appeared before with a known last name, and 13281 first names collected from various dictionaries.	F-scores of 73.66% and 68.08% on Spanish and Dutch CoNLL 2002 datasets, respectively.

Carreras et al. (2002) [15]	Binary AdaBoost classifiers	Included capitalization, trigger words, previous tag prediction, bag of words, gazetteers.	F-scores of 81.39% and 77.05% on Spanish and Dutch CoNLL 2002 datasets, respectively.
Li et al. (2005) [16]	SVM	Experimented with multiple window sizes, features (orthographic, prefixes suffixes, labels, etc.) from neighboring words, weighting neighboring word features according to their position, and class weights to balance positive and negative classes.	F-score of 88.3% on the English CoNLL 2003 data.
Ando and Zhang (2005) [17]	Structural learning [17]	The best classifier for each auxiliary task was selected based on its confidence.	F-scores of 89.31% and 75.27% on English and German, respectively.
Agerri and Rigau (2016) [18]	Semi-supervised system	Included orthography, character of n-grams, lexicons, prefixes, suffixes, bigrams, trigrams, and unsupervised cluster features from the Brown corpus, Clark corpus and k-means clustering of open text using word embeddings.	F-scores of 84.16%, 85.04%, 91.36%, 76.42% on Spanish, Dutch, English, and German CoNLL, respectively.

4 Extracting named entities

Extracting named entities from corpus data allows, on the one hand, to directly retrieve the required data by query, and on the other hand, to test the corpus for containing particular information and to replenish it with documents that include the missing data. The algorithm of extraction of named entities proposed in this paper enables to obtain semantic samples for corpora that do not have semantic data markup. On the other hand, the algorithm has no restriction on semantic types of extracted data, i.e. the semantic type is defined by the keyword in the query.

4.1 Describing algorithm of extracting named entities

The algorithm for extracting named entities is based on the idea of comparing n-grams. The comparison is made within the entire corpus volume, thereby increasing the accuracy of the results.

The extraction process is iterative, the threshold number of iterations specified by the user. The first step presents sampling by the initial search query. The initial search query may be a query on the word form, lemma or phrase, or a search by morphological parameters. A list of bigrams and their frequency is collected across the sample. The bigrams which contain the results are advanced one position to the left or right (set by the user). The resulting list is sorted by frequency of bigrams in order from largest to smallest, to be cut to a predetermined covering index (for example, 95% of all results, this rate being set by the user). This result is used in the second iteration of the algorithm. Each bigram is searched for in the mode of phrasal search in the corpus. Search results are involved in composing a list of trigrams which are advanced one position to the left or right, and their frequency. The resulting list of trigrams is also sorted by frequency in order from largest to smallest, and is cut to a predetermined covering index.

The third and subsequent iterations (until the threshold number of iterations is reached or no match is found as a result of iterating) use the list of n-grams received from the previous iteration. The corpus is searched for each n-gram in the phrasal search mode, and a list of (n + 1)-grams is made up. The resulting list is then cut to a predetermined covering index and compared with the list of n-grams derived from the previous iteration. The comparison accuracy P is set by the user as a percentage. If n-gram frequency is less than P from the quantity of the found (n + 1)-gram, then the n-gram is considered the found named entity, otherwise the extraction proceeds. Thus, the final result will represent a list of the most stable n-grams of different lengths, including search results by the initial search query.

A request to retrieve named entities is an extension of a Q-tuple presented in (1). In addition to the search query, there are added components defining the threshold number of iterations to the left (L) and right (R), the covering index (C), and the accuracy of matching (P). A search example is presented in (1).

$$Q = (Q_1, Q_2, L, R, C, P) \quad (1)$$

4.2 Experiments

Extracting named entities using the algorithm proposed by the authors requires an initial search query which should contain an indicator of a particular named entity. This indicator allows classifying named entities, therefore, the authors chose a set of classes schema.org as the basis for choosing the indicators. From this set of classes, the authors selected the following classes for searching for named entities in the Tatar language corpus: books, restaurants, films, magazines, companies, airports, corporations, languages, technical schools, universi-

ties, schools, shops, museums, and hospitals. Ministries and street names have also been added to this list. Below are some of the results of the experiments conducted by the authors.

Names of ministries

As part of the task of enhancing named entity search a number of experiments have been carried out. One of the most revealing of them was search for names of ministries. The initial search query for the experiment was (2).

$$Q = ((\text{wordform}, \textit{ministrlygy}, \text{“”}, \text{right}, 1, 10, \text{exact}), 7, 0, 95, 80) \quad (2)$$

The result of this query was a list of 50 n-grams containing word form “*ministrlygy*” in the last position. The reference list of names of ministries presented on the Republic of Tatarstan government website [http://prav.tatarstan.ru/tat/ministries.htm] contains 17 items. 12 of 17 items were found in the corpus by means of the algorithm, so the results overlap is 70.6%. 5 items were not found in the corpus for the reasons described in Table 2. The remaining 33 n-grams are different spelling variants of names of ministries.

Table 2. List of unfound names of ministries.

Name	Reason
Urman huҗalygy ministrlygy (Tat) – ministry of forestry	Overlap of the sequence of word forms with the sequence in another name «huҗalygy ministrlygy» (Tat) – ministry of property and «Transport həm yul huҗalygy ministrlygy» (Tat) – ministry of transport and road management
Yashlər eshləre həm sport ministrlygy (Tat) – ministry of youth and sport	Corpus meanings not corresponding to the official name
Transport həm yul huҗalygy ministrlygy (Tat) – ministry of transport and road management	Overlap of the sequence of word forms with the sequence in another name «huҗalygy ministrlygy» (Tat) – ministry of property and «Urman huҗalygy ministrlygy» (Tat) – ministry of forestry
Hezmət, halykny el belən təemin itү həm social yaklau ministrlygy (Tat) - ministry of labour, employment and social protection	Corpus meanings not corresponding to the official name
Ecologia həm tabigy baylyklar ministrlygy (Tat) – ministry of ecology and natural resources	Corpus meanings not corresponding to the official name

Names of streets

Another experiment was concerned with street names search. The search query for this experiment is (3).

$$Q = ((\text{wordform}, \textit{uramy}, \text{“”}, \text{right}, 1, 10, \text{exact}), 7, 0, 95, 80) \quad (3)$$

The result of this query was a list of 600 n-grams containing word form “*uramy*” in the last position. We obtained the following results after manual data evaluation: 432 (72%) n-grams are street names, 72 (12%) n-grams are also street names, but require special character filtering, 96 (16%) n-grams are not street names for various reasons (for example, any sentences containing the word “uramy”; postal addresses and others).

Names of languages

In the next experiment, the authors tried to extract names of languages. The search query for this experiment is presented in (4).

$$Q = ((\text{wordform}, \textit{tel}, \text{“POSS\_3SG,SG”}, \text{right}, 1, 10, \text{exact}), 7, 0, 95, 80) \quad (4)$$

After executing this query, 2310 n-grams were obtained, containing “*tel*” lemma with the morphological properties POSS_3SG and SG in the last position. An estimation of part of the results (a list of 471 n-grams) by an expert showed that in 53.5% of cases (252) n-grams were correct language names. Analysis of the list of n-grams which were incorrectly defined by the algorithm as a name of a language, made it possible to determine additional filtering rules to improve the accuracy of the algorithm. On the basis of the data obtained, the spreading of language names in the corpus of the Tatar language was also constructed (Fig. 1).

Names of restaurants

Another experiment is related to search for names of restaurants. The search query for this experiment is presented in (5).

$$Q = ((\text{wordform}, \textit{restoran}, \text{“POSS\_3SG,SG”}, \text{right}, 1, 10, \text{exact}), 7, 0, 95, 80) \quad (5)$$

The result of this query was a list of 285 n-grams containing “*restoran*” lemma with the morphological properties POSS_3SG and SG in the last position, which in total were found 359 times in the corpus. In this case, in addition to names of restaurants, names of sub-classes of restaurants by their geographical location or national cuisines were obtained.

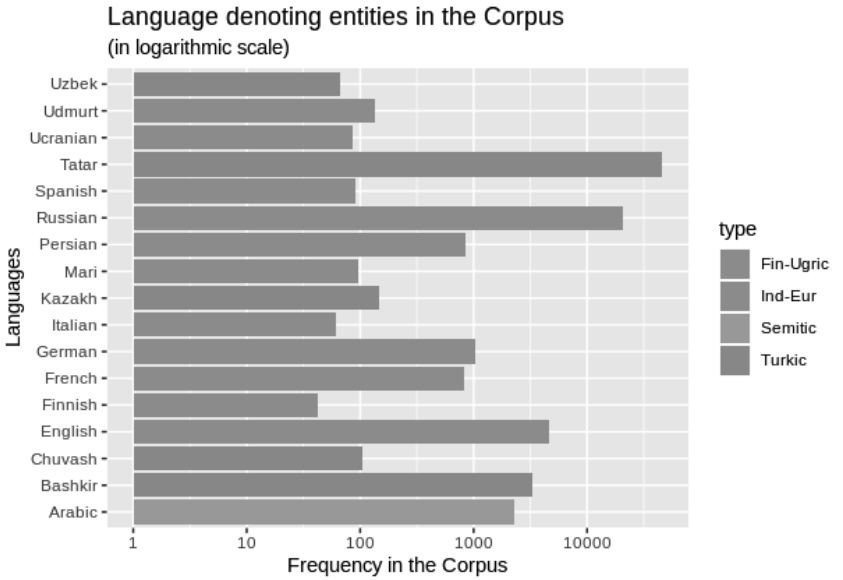


Fig. 1. Language denoting entities in the Corpus.

Thus, 107 (37.68%) found n-grams were correct names of restaurants, their total frequency being 140 (39%). 37 (13.03%) n-grams were the names of subclasses of restaurants, their total frequency being 47 (13.09%). 52 (18.31%) n-grams contained names of restaurants, but they require cleaning from unnecessary parts, while the frequency of the n-grams in the corpus is 2 or less, the total frequency is 54 (15.04%). 45 (15.85%) n-grams contained names of subclasses of restaurants, but they require cleaning from unnecessary parts, while the frequency of n-grams in the corpus is 2 or less, the total frequency is 48 (13.37%). 43 (15.14%) n-grams were not names of restaurants, their total frequency was 65 (18.11%). The list of incorrectly defined n-grams can be reduced by applying additional filtering rules.

Names of corporations

The next experiment was the search for names of corporations. The search query for this experiment is presented in (6).

Q=((wordform,*korporaciya*,"POSS_3SG,SG",right,1,10,exact),7,0,95,80)(6)

As a result of this search query was obtained a list of 138 n-grams containing lemma "*korporaciya*" with morphological properties POSS_3SG and SG in the last position, which were found in the corpus 606 times. Among them, when checked by an expert, 63 (45.65%) n-grams were found, which were correct names of corporations, their total frequency being 178 (29.37%). 27 (19.57%) n-grams contained names of corporations, but require additional cleaning; the total frequency of these n-grams was 29 (4.79%). Among the results, 15 (10.87%) n-grams were singled out, which were non-full names of corporations, their total frequency being 58 (9.57%). 30 (21.74%) n-grams were names of subclasses of corporations by industry, geography, government participation; such n-grams were found in the corpus 336 times (55.45%). 3 (2.17%) n-grams were not names of corporations, their total frequency being 5 (0.83%).

Comparison of results

For different classes of named entities, the algorithm shows different results. The results presented in this article are shown in Table 3.

Table 3. Experiments results.

Class of named entity	Correct	Require filtering	Require expansion	Correct names of subclasses	Names of subclasses that require filtering	Incorrect	Total
Names of ministries	100%	0%	0%	0%	0%	0%	50
Street names	72%	12%	0%	0%	0%	16%	600
Language names	53.5%	0%	0%	0%	0%	46.5%	471 (2310)
Restaurant names	37.7%	18.3%	0%	13%	15.9%	15.1%	285
Corporation names	45.7%	19.6%	10.9%	21.7%	0%	2.2%	138

4.3 Temporal and qualitative indicators of implementing a query for extracting named entities

The experiments showed that the time of implementing a query for extracting named entities depends on the number of found items and bigrams by the initial search query, and on indexes of covering and the accuracy of comparison. All

the experiments were executed on machine with following characteristics: 4 core Intel Core i7 2600 (2,6GHz), 16GB RAM (4x4GB, 1333Hz), SSD 120GB, HDD 3TB (3x1TB, RAID 0). On the test machine Ubuntu Server 14.04 LTS was running. Table 4 shows the timing indicators of search implementation. Algorithm tests revealed dependence of the quality of the results on the number of results found in the first step of the algorithm. This is due to the fact that a smaller number of results increase the actual data coverage and the data which the algorithm works with may initially include particular cases. More results in the first step suggest that at the first cutting of the bigram list, only those will remain that will be included in the final list of the extracted named entities. Thus it is only needed to find the left or the right border for this list.

Table 4. Temporal indexes of implementing searches for extraction of named entities

Search query	Quantity of found items	Quantity of found bigrams	Time elapsed
Q = ((wordform, ministrlygy, “”, right, 1, 10, exact), 7, 0, 97, 80)	27746	68	127.37 sec.
Q = ((wordform, uramy, “”, right, 1, 10, exact), 3, 0, 95, 80)	9592	600	848.07 sec.

5 Conclusion

The algorithm for named entity recognition proposed by the authors in this article shows different results, depending on the type of named entities. The presented results demonstrate correctness of recognition from 37.7% to 100%.

In addition to the main task of named entity recognition, the algorithm is applicable for solving the problem of recognition of names of subclasses of named entities. This feature can be applied to solve additional problems, such as text classification, definition of the subject of texts and other text mining tasks.

Analysis of the results obtained during the experiments show that to improve the accuracy and correctness of the algorithm, its fine tuning, building extended dictionaries for named entity recognition, and additional post-processing of results are necessary.

References

1. Nevzorova O., Mukhamedshin D., Gataullin R. Developing Corpus Management System: Architecture of System and Database. Proceedings of the 2017 International Conference on Information and Knowledge Engineering. CSREA Press, United States of America, pp. 108-112 (2017).

2. Aibaidulla Y., Lua K.T. The development of tagged Uyghur corpus. Proceedings of PACLIC17, pp. 1–3 (2003).
3. Nevzorova, O., Mukhamedshin, D., Kurmanbakiev, M. Semantic aspects of metadata representation in corpus manager system. Open Semantic Technologies for Intelligent Systems (OSTIS-2016), pp. 371–376 (2016).
4. Suleymanov, D., Nevzorova, O., Gatiatullin, A., Gilmullin, R., Hakimov, B. National corpus of the Tatar language "Tugan Tel": grammatical annotation and implementation. Proc. Soc. Behav. Sci. 95, pp. 68–74 (2013).
5. Baldwin B., Carpenter B. LingPipe, http://alias-i.com/lingpipe, last accessed 2018/10/12.
6. Bontcheva K., Dimitrov M., Maynard D., Tablan V., Cunningham H. Shallow methods for named entity coreference resolution. Chaınes de références et résolveurs d'anaphores, workshop TALN. (2002)
7. Zaanen M., Molla D. A named entity recogniser for question answering. Proceedings PACLING (2007)
8. Segura Bedmar I., Mart´ınez P., Herrero Zazo M. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics (2013).
9. Collins M., Singer Y. Unsupervised models for named entity classification. 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (1999).
10. Etzioni O., Cafarella M., Downey D., Popescu A.-M., Shaked T., Soderland S., Weld D., Yates A. Unsupervised named-entity extraction from the web: An experimental study. Artificial intelligence, 165(1), pp. 91–134 (2005).
11. Chinchor N., Robinson P. Muc-7 named entity task definition. In Proceedings of the 7th Conference on Message Understanding, 29 (1997).
12. Pradhan S., Moschitti A., Xue N., Tou Ng H., Bjorkelund A., Uryupina O., Zhang Y., Zhong Z. Towards robust linguistic analysis using ontonotes. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pp. 143–152 (2013).
13. Zhou G., Su J. Named entity recognition using an hmm-based chunk tagger. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. Pp. 473–480 (2002).
14. Malouf R. Markov models for language-independent named entity recognition. Proceedings of the 6th conference on natural language learning, 31 (2002).
15. Carreras X., Marquez L., Padro L. 2002. Named entity extraction using adaboost. Proceedings of the 6th conference on natural language learning, 31 (2002).
16. Li Y., Bontcheva K., Cunningham H. Svm based learning system for information extraction. Deterministic and statistical methods in machine learning. Springer. Pp. 319–339 (2005).
17. Ando R.K., Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research, 6 (Nov), pp. 1817–1853. (2005).
18. Agerri R., Rigau G. Robust multilingual named entity recognition with shallow semisupervised features. Artificial Intelligence, 238, pp. 63–82 (2016).

Logical-ontological approach to coreference resolution

Elena Sidorova[1], Natalya Garanina[1], Irina Kononenko[1] and Alexey Sery[1]

[1] A.P. Ershov Institute of Informatics Systems SB RAS, 6, Acad. Lavrentjev pr., Novosibirsk 630090, Russia
{lsidorova,garanina,alexey.seryj}@iis.nsk.su
irina_k@cn.ru

Abstract. We suggest a logical-ontological approach to the coreference resolution in the process of text analysis and information extraction. Our approach solves the problem of comparing objects found in the text – instances of ontology classes — using the evaluation of the similarity of attributes and relations of objects. In object comparison, we take into account the discourse factors associated with the text and the extra-textual characteristics presented in the ontology of the subject domain. Particularly, we consider polyadic relations which may represent the situations found in the text (events, processes, actions). We propose the ontological interpretation of polyadic relations as classes with single-valued object properties. For coreference resolution we use information about objects and their relations. We propose the corresponding measures for evaluating the semantic similarity of the participant objects in the relations.

Keywords: ontology population, text analysis, information extraction, coreference resolution, referential factors, polyadic relations.

1 Introduction

Identification of referential relations in discourse is one of the most vital but difficult for modeling problems of automatic text analysis. Reference is a relation between some text unit (language expression) and non-linguistic object, which is called a referent. Correct interpretation of an utterance in the text under analysis involves identification of the object mention referent, i.e. reference resolution. There is a range of language means to mention certain referent in the text, and a

speaker (text author) makes choice between two opposite types of language expressions: full noun phrases (proper names and descriptions) and reduced means of reference (pronouns and anaphoric zeroes). Processing expressions of the first type requires direct comparison of extracted objects. In the second case, an anaphoric relation of the reduced expression to antecedent expression is detected with respect to a number of text-structure, syntactic, semantic and pragmatic conditions.

The anaphora and coreference resolution is an important task within the framework of automatic discourse analysis: machine translation, text summarization and information extraction. The latter can be performed by natural language processing in which certain types of information must be recognized and extracted from the text (named entities recognition and fact extraction tasks, in particular). We consider the coreference resolution within the framework of information extraction for ontology population. In this framework, an ontology is used to represent the results of information extraction, and knowledge presented in the ontology helps to solve specific information extraction tasks.

Solving the task of automatic ontology population involves addition of information to the ontology repository. In [1] we consider mentions of simple entities and propose an approach to their coreference resolution in the process of information extraction for ontology population. An ontology structure allows to take into account implicit information in the input text due to detecting relations between objects. In this paper, we suggest coreference resolution for new objects with a complex structure including situations (events, actions, processes), which are represented by polyadic relations in an ontology. These situations extend the domain knowledge used for solving coreference resolution problem. The new knowledge improves the quality of coreference resolution.

In Section 2 we give a brief review of modern trends in the coreference problem definition and the present research. In Section 3 we describe our basic approach to ontology-based information extraction with formal definitions of and ontology and polyadic relations. Section 4 presents ontological factors relevant for coreference resolution illustrated by text examples and revises the similarity measure of objects. In Section 5 we consider features of experiments in our approach. We conclude with the base characteristics and advantages of the proposed approach and outline the directions for future research.

2 Coreference in Information Extraction Tasks

We observe several classification aspects of problems related to the reference identification.

- First aspect is the way of presenting references in the text: full lexical expressions (noun phrases – proper names, descriptions, descriptions combined with proper names) or reduced expressions using anaphoric means (pronouns, determiners) or anaphoric zero. In the first case, for noun phrases based on proper names, the problem is detecting identical references to named entities. In the second case, the problem is identification of the antecedent, i.e. anaphora resolution [2, 3].
- Second aspect is the type of the referenced object: referential identity of entities or situations (events).
- Third aspect is the search area and type of context: the context of a single document (simple and complex sentences or chains of sentences in one text) opposes to cross-document analysis, in which references to the same object are looked for in the corpus or document flow.

The traditional problem of anaphora and coreference resolution within a coherent text remains to be relevant. Many early and modern researches solve the problem using linguistic methods based on rules and methods of machine learning. R. Mitkov's reviews [4, 5] and later [6, 7] consider the basic approaches to this problem. Recently, there has been a growing interest in solving the problem in a broader perspective: not only entities but also events or situations have been considered [8 – 12]. A cross-document reference analysis that is an important approach for populating knowledge bases and ontologies is used for the problem as well [8, 13 – 15]. The complexity of the problem of coreference resolution requires an integrated approach, involving both knowledge about the structure of the text (the level of discourse) and knowledge about the subject area, which are determined by the classes of entities in a specific ontology and their ontological structure (ontological level). In [16] the authors consciously abstract away from the discourse factors of coreference in order to investigate the role of subject knowledge. Discourse features represent the structural and textual properties of mentions (similarity of sub-chains, position, distance), grammatical and lexical features. Obviously, new tasks require a revision of the role of discourse features in comparison with ontological ones. Thus, cross-document analysis does not consider pronominal anaphora and hardly takes into account such discourse factors as the order of appearance of mentions in the text, and the distance (linear or rhetorical).

Theories of discourse analysis distinguish several types of discourse connectivity: referential (identity of participants), spatial, temporal and event-triggered ones [17]. In applied research, there are two approaches to understanding the coreference of events. In the first approach, two mentions of an event are considered coreferent if they are characterized by the same set of properties (such as time or place of the event) and the same set of participants [9 – 11]. In

the second approach, only the referential identity of participants is considered for referential identity of events [3]). In [12] a broader set of referential relations between two mentions of events is considered: complete coreference, subevents for vertices of the parent and child layers, subevents for a descendant vertex of a single layer.

We consider the problem of information extraction as a task of detecting all references to objects of a given domain: entities and situations (events, states, actions, processes). In the ontology population task, the found objects should be represented as instances of concepts and relations of the ontology. It is necessary to establish referential relations between all instances found in the process of text analysis and instances of the ontology information content (which does not exclude the possibility of adding new instances to the ontology).

3 The Model of Information Extraction

Consider the environment in which our approach to coreference resolution is being developed. Fig. 1 shows the general scheme of the information extraction system (IE-system) with the emphasized module of coreference resolution.

The input of our IE-system comprises: the ontology of a subject domain, the ontology population rules and the results of preliminary text processing including the terminological, thematic, and segment coverings of an input text.

A terminological covering is the result of lexical text analysis which extracts terms of a subject domain from a text and forms lexical objects using semantic vocabularies. A segment text covering is a division of the text into formal fragments (clauses, sentences, paragraphs, headlines, etc.) and genre fragments (document title, annotation, glossary, etc.). A thematic covering selects text fragments of a particular topic. A construction of a thematic covering is based on the thematic classification methods.

The module of information extraction constructs objects representing instances of concepts and relations of the domain ontology from the lexical objects [18]. This module uses the ontology population rules which are automatically generated from fact schemes. The fact schemes are formulated by experts taking into account the ontology and language of a subject domain. These fact schemes constrain morphological, syntactic, structural, lexical, and semantic characteristics of the objects.

The coreference resolution module [19] runs in parallel with the information extraction module. This module forms hypotheses about coreference relations, and calculates their weights using various factors discussed below.

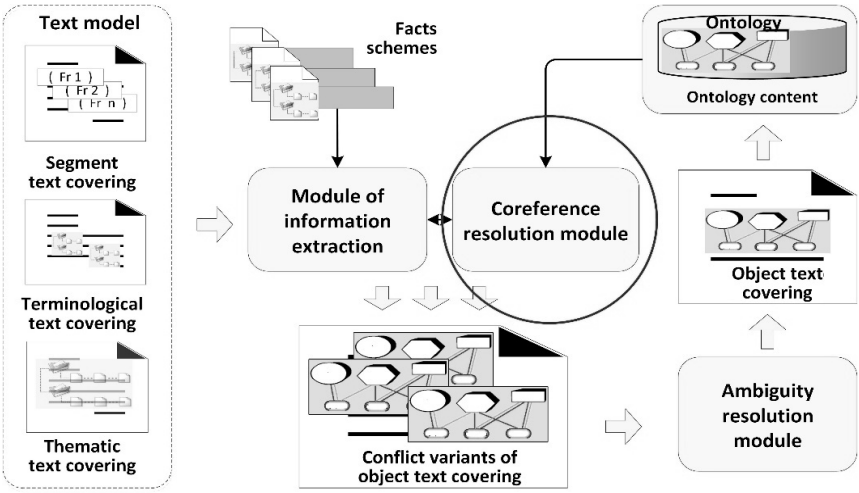


Fig. 1. The scheme of the system of information extraction and ontology population.

The ambiguity resolution module resolves all types of conflicts which are the result of various interpretations of the input text — different object text coverings for the same text fragment. This module chooses the most informative variant from the set of possible interpretations (the variant with the highest weight) [20].

The result of the work of our IE-system is the population of ontology content by instances of concepts and relations of the subject domain found in the input text.

3.1 The Ontology of a Subject Domain

An ontology O of a subject domain includes the following elements:

- a finite nonempty set C_O of *classes* for representing the concepts of the subject domain,
- a finite set D_O of *data domains*, and
- a finite set of *attributes* with names in $Atr_O = Dat_O \cup Rel_O$, each of which has values in some data domain from D_O (*data attributes* or *datatype properties* in Dat_O) or has values as instances of some classes (*object attributes* or *object properties* in Rel_O, which model binary relations).

Each class $c \in C_O$ is defined by the set of its attributes: $c = (Dat_c, Rel_c)$, where every data attribute $\alpha \in Dat_c \subseteq Dat_O$ has the domain $d_\alpha \in D_O$ with values in V_{d_α} and every object attribute $\rho \in Rel_c \subseteq Rel_O$ has values from the subset $C_\rho \subseteq C_O$.

The set of all class attributes is denoted by $Atr_c = Dat_c \cup Rel_c$. We consider an ontology without data and class synonyms, i.e. $\forall\ \alpha_1, \alpha_2 \in Dat_O$: $d_{a1} \neq d_{a2}$ and $\forall\ c_1, c_2 \in C_O$: $Atr_{c1} \neq Atr_{c2}$.

We denote the class of an attribute γ by c^γ and the set of its values by D^γ. A set of attributes of every class must include the nonempty set of *key attributes* Atr_c^K. The key attributes can either be data or object attributes. These attributes guarantee unambiguous definition and uniqueness of the class instances.

A tuple $a = (c_a, Dat_a, Rel_a)$ is *an instance of the class* $c_a = (Dat_{c_a}, \mathrm{Rel}_{c_a})$ $(a \in c_a)$ iff every data attribute $\alpha_a \in Dat_a$ has a name $\alpha \in Dat_{c_a}$ with the values V_{α_a} from V_{d_α} and every object attribute $\rho_a \in \mathrm{Rel}_a$ has a name $\rho \in \mathrm{Rel}_{c_a}$ with the values V_{ρ_a} as instances of the classes from C_ρ.

We use the standard *class inheritance relation*: the class c_2 *is a subclass* of the class c_1 ($c_1 < c_2$) iff $\forall\ a \in c_2$: $a \in c_1$.

The information content IC_O of the ontology O is a set of instances of the classes from O. *The ontology population problem* is to compute information content for a given ontology from the given input data.

3.2 Polyadic Relations

The notion of polyadic relation is not considered in the classical ontology theory. For example, the OWL – the standard ontology description language – has no language constructions for polyadic relations, only binary relations (Object Property) are available. On the other hand, polyadic relations frequently arise in the tasks of extracting information from texts, because they can describe the propositional content of a statement that represents an extra-linguistic situation, or state of affairs (event, action, process, etc.).

To overcome these shortcomings, we model *polyadic relations* (or just relations) by ontology classes with constraints on the set of attributes. First, relations classes have to include at least two object properties. Second, every object property of a relation has to be a key attribute. A polyadic relation may also contain datatype properties without special constraints.

Due to this definition, a polyadic relation is naturally represented by the set of binary relations. And vice versa, a binary relation can be represented by the polyadic relation with two object properties as a special case of polyadic relations.

In text processing, we consider polyadic relations correspond to descriptions of situations (actions, processes) and other objects with complex structure. The following Table 1 gives some examples of polyadic relations extracted from texts.

These examples relate to the automated control systems subject domain that includes such relation classes as *Action*, *Process*, *Function*, *Control*, *Movement*, *Change_of_state*, etc. Object properties of relation classes correspond to the hierarchy of semantic roles. The semantic role is a generalization of the functions of a participant in a range of situations denoted by a group of predicates, and hence the types of corresponding situations.

Table 1. Examples of polyadic relations

S1 Action	**Type: information_transfer** **Sender: X** **Recipient: Y** **Message: Z** **Content: null**	The system (Y) receives commands (Z) from the operator (X)
S2 Process	**Agent: X2** **Type: processing** **Message: Z**	The command (Z) is entered by the operator (X2) through the remote operator console

3.3 The Coreference Resolution Problem

The information content of a text consists of a set of instances of ontology classes and relations found in the text, which are provided with additional information. We define a set A of *information-text objects (i-objects)* retrieved from input data and corresponding to ontology instances. Every i-object $a \in A$ has the form $(c_a, Dat_a, Rel_a, G_a, P_a)$, where

— $c_a \in C_O$ is the ontology class;

Dat_a is the set of data attributes $\alpha_a = (\alpha, V_{\alpha_a})$, where

- $\alpha \in Dat_{c_a}$ is the attribute name, and V_{α_a} is the set of values $v \in d_\alpha$;

Rel_a is the set of object attributes $\rho_a = (\rho, V_{\rho_a})$, where

- $\rho \in \mathrm{Rel}_{c_a}$ is the attribute name, and V_{ρ_a} is the set of i-objects of a class $c_{\rho_a} \in C_{\rho_a}$;

— G_a is the grammar information (morphological and syntactic features based on grammar features of lexical object);

— P_a is the structural information (a set of positions in the input data and the formal segments).

The attribute γ of the i-object a is *filled* if $V_{\gamma_a} \neq \varnothing$. We denote by $Atr_a = Dat_a \cup Rel_a$ the set of all attributes. Each i-object corresponds to some ontology instance in a natural way as follows. Let $a = (c_a, Dat_a, Rel_a, G_a, P_a)$ be an i-object, then its corresponding ontology instance is $a' = (c_a, Dat_a, Rel_a)$, and every $\alpha \in Dat_{a'}$ has value(s) in V_{α_a} and every $\rho \in Rel_{a'}$ has values in V_{ρ_a}.

We assume that i-objects a and b are *possible coreferents* $a \approx b$ (*candidates for coreference*) iff their classes are transitively related by the class inheritance

relation and the set of values of all filled key attributes of one i-object is included in the set of values of the corresponding key attributes of the other i-object.

The coreference resolution problem is to detect if given candidates for coreference correspond to the same ontology instance.

4 Referential Factors

In previous papers [19], we considered two types of factors that affect the evaluation of the measure of the coreferential similarity of two objects. First, discourse factors (local textual and contextual) are determined by the language means used to represent the objects in the text and by their location in the text structure. Second, semantic factors determine the similarities of objects with respect to their ontological structure and relations.

In our approach, we distinguish logical-ontological factors for considering a set of associated relations between objects. For these factors we use the properties of relations specified in the ontology.

All these factors are used to evaluate similarity of objects mentioned in the text. For each factor, we define a similarity measure. This measure corresponds to the degree of strength of the coreferent relation between the i-objects *a* and *b* with respect to the factor, without taking into account other factors.

4.1 The Coreferential Conflict and the Similarity Measure

We define *coreferential conflict* as a case when two non-coreferent i-objects *a* and *b* are possible coreferents of the third i-object *c*: $a \leftrightsquigarrow^{c} b \Leftrightarrow (a \approx c) \wedge (b \approx c) \wedge \neg(a \approx b)$.

To determine which of these i-objects are actually coreferent, we use the measure of coreference similarity of i-objects. This measure for i-objects *a* and *b* is denoted as *cs(a,b)*. If the non-coreferential i-objects *a* and *b* are possible coreferents for the i-object *c*, we say that *the coreferential conflict is resolved to a* iff *cs(a,c) > cs (b,c)*, i.e. the i-object *a* is more similar to i-object *c*, then i-object *b*.

The integral measure of similarity *cs(a,b)* is calculated as an Euclidean measure of similarity based on four measures – semantic *S(a,b),* context *C(a,b)*, position *P(a,b)* and grammar *G(a,b)*.

$$cs(a,b) = \frac{1}{2}\sqrt{(1-S(a,b)^{2} + (1-C(a,b)^{2} + (1-P(a,b)^{2} + (1-G(a,b)^{2}} \quad (1)$$

The context similarity measure C(a,b) takes into account the information connectivity of i-objects in a given text. This measure depends on the number of i-objects which directly or indirectly use a) attribute values from both *a* and *b*, and b) attribute values borrowed by *a* from *b*, and by *b* from *a*, for the evaluation of their own attributes.

The position similarity measure P(a,b) takes into account variants of location of i-objects in an input text. This measure depends on the number of segments, number of possible candidates in the conflict, and number of lexemes placed between the positions of *a* and *b*.

The grammar similarity measure G(a,b) is based on the standard linguistic features such as gender, number, person, etc.

The semantic similarity measure S(a,b) determines the degree of proximity of the corresponding attribute sets Atr_a and Atr_b. Comparing these two sets takes into account both the similarity of the values of their constituent elements and additional characteristics based on the ontological properties of attributes, including the inheritance of classes and data attributes, intersection, union, composition, refinement, inversion, inclusion, closure, transitivity and symmetry.

In [1] we consider 11 types of similarities. Below we expand this set with similarities using polyadic relations. Initially, *S(a,b)* was determined by formula (2), where $Sim_b^a = \{(\alpha_a, \beta_b) \mid sim(\alpha_a, \beta_b) \neq 0\}$:

$$S(a,b) = \frac{1}{|\, Sim_b^a \,|} \sum_{(\alpha_a, \beta_b) \in Sim_b^a} sim(\alpha_a, \beta_b) \qquad (2)$$

Here, under the sign of the sum, all kinds of similarities of the attributes of the objects *a* and *b* are collected. Practical considerations and experimental data revealed particular cases in which basic formula (2) is inexact and instable with respect to adding new attribute comparison characteristics: i-objects that have a large set of comparable but actually not similar attributes can turn out to be close with each other due to just taking into account that the similarity of attributes that is greater than zero. It is worth noting that such cases are very rare due to the definition of coreference and the formulation of the problem of extracting i-objects. The second disadvantage of formula (2) is expressed by the fact that adding new terms to the sum can decrease the total value. But one should expect that positive additional information about the proximity of attributes have to always increase the similarity of the corresponding i-objects. These additional characteristics are based on the ontological properties of attributes, including, in particular, composition, transitivity, refinement, etc., and specials properties

of polyadic relations described below. In view of the above, it was proposed to convert formula (2) to a formula of the following form:

$$S(a,b) = S^{EQ} + (1 - S^{EQ}) \cdot S^{\Delta} \tag{3}$$

The value $S^{EQ} \in [0;1]$ corresponds to the similarity of the values of the corresponding attributes of the objects *a* and *b* without taking into account the additional characteristics, and $S^{\Delta} \in [0;1)$ – the additional information provided by these characteristics.

S^{EQ} is calculated by formula (4), similar to formula (2), where the set of pairs of similar attributes Sim_b^a is replaced by the set of pairs of comparable attributes $Comp_b^a = \{(\alpha_a, \beta_b) \mid \alpha_a \in Atr_a, \beta_b \in Atr_b, \alpha = \beta\}$.

$$S^{EQ} = \frac{1}{|Comp_b^a|} \sum_{(\alpha_a, \beta_b) \in Comp_b^a} sim(\alpha_a, \beta_b) \tag{4}$$

Only measures of standard similarity of attributes by values stand under the sign of the sum in the formula (4) [19].

Let the total amount of additional information about the attributes of objects *a* and *b* be

$$I = \sum_{\gamma_a \in Attr_a, \delta_b \in Attr_b} sim^{\Delta}(\gamma_a, \delta_b) \tag{5}$$

Here the symbol Δ denotes additional properties of attributes, such as transitivity, composition, etc. It is obvious that *I* can take any positive values. Hence, in order to get the value of S^{Δ} varying from 0 to 1, we need a monotonic transformation defined everywhere on the positive semi-axis. Using *I*, we evaluate the additional similarity of the i-objects *a* and *b*. Really, we determine the value of the probability of this similarity S^{Δ}:

$$S^{\Delta} = \frac{I}{1 + I} \tag{6}$$

We can see from formulas (3), (5) and (6) that

— $S(a,b) = 1 \Leftrightarrow S^{EQ} = 1$,

— $S^{\Delta} \in [0;1)$, and

— $S(a,b) > S^{EQ} \Leftrightarrow S^{EQ} < 1 \wedge S^{\Delta} > 0$.

In other words, when objects have incomplete similarity in the values of comparable attributes, and the additional information is available, the degree of similarity *S* is always greater than S^{EQ}, but full match is achieved only under

the condition that the values of all comparable attributes are the same taking coreference into account.

4.2 Relations Factor

For evaluating similarity we consider polyadic relations in the following two aspects.

First, comparing polyadic relation instances for identification coreference between them.

Example 1. *When the bottle reaches a certain position, (the sensor^{X} communicates with the conveyor $^{Y})^{S1}$ to inform it that it should stop. For this purpose (the sensor^{X} sends a signal Stop^{Z} to the receiving device of the conveyor$^{Y})^{S2}$*

In this example, we can distinguish two possible coreferent instances of polyadic relations *S1* and *S2*:

- S1: Contact (Originator: X, Recipient: Y)
- S2: Information_transfer (Originator: X, Recipient: Y, Content: Z)

These instances are similar because their *Originator* and *Recipient* attributes have coreferent values.

Second, using information about polyadic relations for identification coreference between i-objects participating in these relations. For this purpose, pairs of relations are considered that contain similar values (besides the objects themselves being compared). Change the example from the previous version.

Example 2. *(The sensor^{X1} transmits a message^{Z} to the conveyor$^{Y})^{S1}$ to inform it that the bottle has reached a certain position. So, (it^{X2} controls the operation of the conveyor$^{Y})^{S2}$.*

In this example polyadic relations are represented by the following instances:

- S1: Information_transfer (Originator: X1, Recipient: Y, Content: Z)
- S2: Control (Controller: X2, Patent: Y)

We consider the instances X1 and X2 are similar because S1 and S2 have a similar value Y. Note that in the last example the relations of different classes with different sets of object attributes are compared because we allow the comparison of arbitrary relations.

We define the following formal ontological properties for object attributes. They are used for definition of object similarity measures that take into account polyadic relations. We borrow some concepts of relational algebra. We denote the set of all polyadic relations of the ontology O by S_O.

Definition 1. Let $\rho, \rho', \rho'' \in Rel_O$.

– The attributes ρ, ρ′ are in the projection relation $\rho=_{\pi}\rho'$ iff $C_{\rho}, C_{\rho'} \subseteq S_O$ and $\exists \bar{A}=(\gamma_1,\ldots,\gamma_m), \bar{A}'=(\gamma'_1,\ldots,\gamma'_m)$:$\forall a\in c\in C_{\rho} \exists a' \in C_{\rho'}$: $\pi_{\bar{A}}(a) = \pi_{\bar{A}'}(a')$, i.e. $V_{\gamma ia} = V_{\gamma' ia'}$ (i∈[1..m]), and vice versa, $\forall a'\in c'\in C_{\rho'} \exists a \in C_{\rho}$: $\pi_{\bar{A}'}(a') = \pi_{\bar{A}}(a)$, i.e. the values of the attributes that are in the projection relation are instances of the polyadic relations that contain equal values.

– The attributes ρ, ρ′ and ρ″ are in the natural join relation $\rho=\rho'\bowtie\rho''$ iff C_{ρ}, $C_{\rho'}$, $C_{\rho''}\subseteq S_O$ and $\forall a'\in c'\in C_{\rho'}, \exists a \in c \in C_{\rho}, A \subseteq Atr_a$: $\pi_{Atra'}(a') = \pi_A(a)$, $\forall a''\in c''\in C_{\rho''} \exists a \in c \in C_{\rho}, A \subseteq Atr_a$: $\pi_{Atra''}(a'') = \pi_A(a)$, and $\forall a\in c\in C_{\rho}, b\in Atr_a$: ($\exists a'\in c'\in C_{\rho'}$, $b'\in Atr_{a'}$: $b=b'$)$\vee$ ($\exists a''\in c''\in C_{\rho''}$, $b''\in Atr_{a''}$: $b=b''$), i.e. the instances that are the values of the object attributes ρ′ и ρ″ are complementary different views (projections) on the values of the attribute ρ.

Thus, the projection describes a subset of the common elements of the relation instances. In Example 1, the common projection of instances of the relations *S1* and *S2* is {*X, Y*}. In Example 2, the corresponding projection is the set {*Y, X1, X2*}. The natural join takes into account the presence of a third relation when comparing a pair of relation instances. This relation includes the join of the attributes of these relations. The presence of such a third relation is an evidence of the information included in the first two ones.

The example of the ontological natural join relation is ontological description of the modules of a technological complex that execute the similar tasks. Each module is represented by a relation, including instances of the tasks: S_{Mi} *(*w_1*, ...,* w_n*)*. The complex performs the whole set of tasks, which is the result of the natural join of the tasks executed by the modules: $\cup w_{ij}$, $w_{ij}\in S_{Mi}$.

For those cases when properties of attributes in Definition 1 cannot be derived from the ontology description, there is a need to check the necessary conditions of the presence of the properties. The following proposition formulates these conditions in a constructive way. We denote the necessary condition of a property *x* by $\mathcal{N}^x$.

Proposition 1. Let $\rho, \rho', \rho'' \in Rel_O$.
— $\rho=_{\pi}\rho' \Rightarrow \mathcal{N}^{\pi} = (C_{\rho}\cap^i C_{\rho'} \neq \emptyset)$;
— $\rho=\rho'\bowtie\rho'' \Rightarrow \mathcal{N}^{\bowtie} = (C_{\rho'}\cup^i C_{\rho''}\subseteq {}^i C_{\rho})$.

Here, the superscript *i* in the set operations means that we make the operation over the elements of the sets and over their parental classes and subclasses in the class hierarchy. The proof follows from Definition 1.

Taking into account Definition 1, we define the projection and natural join based similarities of the attributes. We also define the class similarity. In the following definition, the superscript *r* in comparison operations and calculation of the power of sets means that the operations consider the elements of the sets and their possible coreferents.

Definition 2. For *i*-objects *a* and *b* with $a \approx b$ and $c_a \leq c_b$, we compute *the power of the class similarity* as $sim^c(c_a, c_b) = |c_b|/|c_a|$, where $|c_x|$ is the number of subclasses of the class *x* including *x* itself.

Definition 3. For i-objects *a* and *b*, we consider object relation $\rho \in Rel_a$ and $\xi \in Rel_b$ with $\rho, \xi \in S_O$ is

– *projectionally similar* $\rho \sim_\pi \xi$, iff $\rho =_\pi \xi \vee \mathcal{N}^\pi$ and $S^\pi = \cup_{x \in V\rho a} \{X \subseteq Atr_x \mid \exists y \in V_{\xi b}, Y \subseteq Atr_y : \pi_X(x) =^r \pi_Y(y)\} \neq \emptyset$. The power of the projection similarity is $sim^\pi(\rho, \xi) = ½|S^\pi|(c(V_{\rho a})^{-1} + c(V_{\xi b})^{-1})$, where $c(V_\mu) = \sum_{z \in V\mu} \sum_{\gamma \in Atrz} |V_\gamma|^r$.

– *joinly similar* $\rho \sim_{\bowtie} \xi$, iff $\exists\, \mu: \mu = \rho \bowtie \xi \vee \mathcal{N}^\bowtie$ and $S^\bowtie = \{(x, y) \mid x \in V_{\rho a}, y \in V_{\xi b}, \exists\, z \in C^\mu, Z_x \subseteq Atr_z, Z_y \subseteq Atr_z: Atr_z \subseteq Z_x \cup Z_y, \pi_{Atrx}(x) =^r \pi_{Zx}(z)$ and $\pi_{Atry}(y) =^r \pi_{Zx}(z)\} \neq \emptyset$. The power of the join similarity is $sim^\bowtie(\rho, \xi) = ½|S^\bowtie|((|V_{\rho a}|^r)^{-1} + (|V_{\xi b}|^r)^{-1})$.

Thus, we can take into account the power of sim^c, sim^π and $sim^\bowtie$ of the projection and join similarity in the semantic similarity measure along with the other factors in formula (5). This allows us to take the context into account more accurately, improving the quality of information extraction.

5 Characteristic of Experimental Study

The proposed approach to resolving coreference is based on the properties of the domain concepts presented formally. Testing its implementation requires for a formally presented ontology of a subject domain, as well as text corpus annotated in accordance with the ontology. Typed coreferential relations also have to be annotated.

There exist coreferentially annotated corpora for English (MUC) and a number of other languages (Catalan, Dutch, English, German, Italian, Spanish, Czech, Chinese and Arabic). The first open corpus for Russian is RuCor (available at http://rucoref.maimbava.net/) that represents anaphorical and coreferential relations and morphological annotation. RuCor contains about 200 texts of different genres (primarily news, essays, and fiction) that do not correspond to any special subject domain [21]. The lack of appropriate datasets with deep layers of annotation is the obstacle to the study of complex cases of coreference.

Hence, for evaluation of our approach we form a corpus of examples with a complex type of coreference, which can be resolved on the basis of ontology. Several examples are selected for each type of ontological relation. The total volume of the corpus is about 50 text fragments taken from texts of technical documentation and encyclopedias. These fragments represent specifications of requirements from the subject domain of automated control systems. Each example is annotated by coreference relations with types based on ontological properties.

We consider such annotation of coreference information necessary for further linguistic research. Extending the capabilities of automatic analyzers with computational similarity models based on ontological properties improves the quality of coreference resolution. Thus, for the examples found, the use of logical-ontological measures allows to increase the measure of similarity of the "correct" variant by 0.05-0.1 (5-10%).

6 Conclusion

In the papers on the topic of coreference resolution, we proposed a formal statement of the problem and mathematically-strict definitions of the notions of coreference, coreferential conflict and ontological properties used to resolve the coreference. This is an important contribution to ensure the correct operation and improve the quality of the coreference resolution algorithms.

The main features of the proposed approach to coreference resolution are:

1. shift of the emphasis from discourse factors to the subject knowledge, primarily to the ontology of the subject domain to be populated through information extraction, disambiguation, and coreference resolution;

2. integration of computational and linguistic models and techniques of text analysis at the phase of semantic processing. Thus, weighted coreferential relations between objects are used for coreference resolution. In this process, the hypothetic coreferential relations are generated by the linguistic model, and the resolution (choice of the best hypothesis) is based on the statistical data;

3. scalability of the solution. Our approach can be enriched with new information extraction rules and referential factors.

The corpus with annotated coreference is necessary for studying different cases of repeated mentions of events that need ontological information about polyadic relations to correctly resolve coreferences. Our future research will focus on general classification of such cases. We plan to develop special case-oriented coreference resolution techniques, particularly, by considering the relevance of ontological properties for the evaluation of similarity of possible coreferents. Taking this into account, we are faced with the problem of defining ontology formal properties that provide a better solution to the tasks of extracting information from the text and, in particular, the resolution of the coreference.

Acknowledgement.

The study was supported by the Russian Foundation for Basic Research, project 17-07-01600.

References

1. Garanina, N., Sidorova, E., Kononenko, I., Gorlatch, S.: Using Multiple Semantic Measures For Coreference Resolution. Ontology Population. International Journal of Computing 16(3), 166–176 (2017).
2. Dimitrov, M., Bontcheva, K., Cunningham, H., Maynard, D.: A Light-weight Approach to Coreference Resolution for Named Entities in Text. In: Branco, A., McEnery, T., Mitkov, R. (eds.) Anaphora Processing: Linguistic, Cognitive and Computational Modelling, vol. 263, pp. 97-112. John Benjamins Publ., (2005).
3. Sobha, L.: Anaphora Resolution Using Named Entity and Ontology. In: Johansson, C. (ed.) Proceedings of the Second Workshop on Anaphora Resolution (WAR II), NEALT Proceedings Series, vol. 2, pp.91-96 (2008).
4. Mitkov, R.; Anaphora resolution: the state of the art. In: Working paper based on the COLING'98/ACL'98 tutorial on anaphora resolution. Wolverhampton (1999).
5. Mitkov, R.: Anaphora resolution. In: Mitkov, R. (ed.) The Oxford handbook of computational linguistics, ch.14, pp. 266-283. Oxford university press, N.Y. (2003), https://pdfs.semanticscholar.org/e782/00b1e3ba2a72de1ca9b9b2c5efa775151bfa.pdf, last accessed 2018/10/04.
6. Elango, P.: Coreference Resolution: A Survey: Technical Report. UW-Madison (2006), https://ccc.inaoep.mx/~villasen/index_archivos/cursoTATII/EntidadesNombradas/Elango-SurveyCoreferenceResolution.pdf, last accessed 2018/04/01.
7. Prokofyev, R., Tonon, A., Luggen, M., Vouilloz, L., Difallah, D.E., Cudr´e-Mauroux, P.: SANAPHOR: Ontology-Based Coreference Resolution. In: 14th International Semantic Web Conference, part I, LNCS, vol. 9366, pp. 458-473. Springer, Cham (2015).
8. Lee, H., Recasens, M., Chang, A., Surdeanu, M., Jurafsky D.: Joint Entity and Event Coreference Resolution across Documents. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language, EMNLP-CoNLL 2012, pp. 489–500 (2012).
9. Cybulska, A., Vossen, P.: "Bag of Events" Approach to Event Coreference Resolution. Supervised Classification of Event Templates. International Journal of Computational Linguistics and Applications 6(2), 11-27 (2015).
10. Borgo, S., Bozzato, L., Aprosio, A.P., Rospocher, M., Serafini L.: On Coreferring Text-extracted Event Descriptions with the aid of Ontological Reasoning. Technical Report (2016), https://arxiv.org/pdf/1612.00227.pdf, last accessed 2018/10/04.
11. Bejan, C.A., Harabagiu, S.: Unsupervised event coreference resolution with rich linguistic features. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp.1412-1422 (2010).
12. Araki, J., Liu, Z., Hovy, E., Mitamura, T.: Detecting Subevent Structure for Event Coreference Resolution. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), pp. 4553–4558 (2014).
13. Mayfield, J., Alexander, D., Dorr, B.J., Eisner, J., Elsayed, T., Finin, T., Fink, C., Freedman, M., Garera, N., McNamee, P., Mohammad, S., Oard, D., Piatko, C., Sayeed, A.B., Syed, Z., Weischedel, R.M., Xu, T., Yarowsky, D.: Cross-Document

Coreference Resolution: A Key Technology for Learning by Reading Association for the Advancement of Artificial Intelligence. In: AAAI Spring Symposium: Learning by Reading and Learning to Read, pp.65-70 (2009).
14. Yatskevich M., Welty C., Murdock J.W. Coreference resolution on RDF Graphs generated from Information Extraction: first results. ISWC'06 Workshop on Web Content Mining with Human Language Technologies (2006).
15. Hladky, D., Ehrlich, C., Efimenko, I., Vorobyov V.: Discover Shadow Groups from the Dark Web. In: Web Intelligence and Security: Advances in Data and Text Mining Techniques for Detecting and Preventing Terrorist Activities on the Web, pp. 67-81 (2010).
16. Suleymanova, E., Trofimov, I.: A method for coreference resolution within information extraction. In: Program Systems: Theory and Applications 1(15), 15–30 (2013). (in Russian)
17. Giv`on T.: Coherence in text, coherence in mind. Pragmatics and cognition 1(2), 171–227 (1993).
18. Garanina, N., Sidorova, E.: Ontology Population as Algebraic Information System Processing Based on Multi-agent Natural Language Text Analysis Algorithms. Programming and Computer Software 41(3), 140–148 (2015).
19. Garanina, N, Sidorova, E., Seryi, A.: Multiagent Approach to Coreference Resolution Based on the Multifactor Similarity in Ontology Population. Programming and Computer Software 44(1), 23–34 (2018).
20. Garanina, N., Sidorova, E., Anureev, I.: Conflict resolution in multi-agent systems with typed relations for ontology population. Programming and Computer Software 42(4), 31–45 (2016).
21. Toldova, S., Roytberg, A., Nedoluzhko, A., Kurzukov, M., Ladygina, A., Vasilyeva, M., Azerkovich, I., Grishina, Y., Sim, G., Ivanova, A., Gorshkov, D.: Evaluating Anaphora and Coreference Resolution for Russian. In: Computational Linguistics and Intellectual Technologies, Proceedings of the International Conference "Dialog 2013", pp. 681–695. Publishing House of the RSUH, Moscow (2013).

Application of KEA for semantically associated structural units search in a corpus and text summarization

E.V. Sokolova

Saint Petersburg State University, Saint Petersburg, Russia
st049868@student.spbu.ru

Abstract. This paper presents results of the research on possible applications of keyphrase extraction algorithm KEA. Although this algorithm is widely used as an effective and universal tool for keyphrase extraction, our study is aimed at exploration of its further adjustments in the tasks of translation equivalents search and for semantic compression, namely, for extractive summarization. To be precise, in our first series of experiments we analyzed the output of KEA based on the text corpus developed from the United Nations documents in order to find semantically associated structural units (possible translation equivalents) among Russian and English keyphrases. The second series of experiments is concerned with using keyphrases automatically extracted by KEA to compose extracts for short stories. In this case we also compiled a corpus of short stories written in (or translated into) Russian and adjusted KEA so that ranked sentences with keyphrases could be used to form previews for the stories.

Keywords: keyphrase extraction, KEA, translation equivalents, summarization.

1 Introduction

Keyphrases have a wide range of practical applications in rather different fields such as document summarization, indexing, information retrieval, library systems, etc. Being structural units themselves, keyphrases convey the most important information about the content of the document. That is why automatic keyphrase extraction is one of the most highly sought tasks to solve today.

There are different approaches to extract keyphrases from a document [1, 2]: statistical (TFxIDF, Chi-square, C-value, Log-Likelihood, etc.), linguistic (in-

cluding different levels of linguistic analysis), machine learning (Naïve Bayes classifier, SVM, etc.) and also hybrid algorithms (KEA).

In this paper we explore further implementations of one of commonly known keypharse extraction algorithms KEA (Keyphrase Extraction Algorithm) in the wide field of Natural Language Processing (NLP) [3]. Therefore, we conducted a series of experiments trying to adjust KEA to the tasks which combine semantic compression and text transformations.

To be precise, in the first experiment we try to find out if KEA is capable of finding semantically related unites, such as translation equivalents, synonyms, hyponyms, etc., for two different languages, namely Russian and English.

The second experiment is devoted to the possibility of using KEA as an intermediate tool for an extractive summarization [4] algorithm. Keyphrases automatically extracted by KEA were used to identify salient sentences in the text.

To mark the borders of our research, it needs to be noted that we are not trying to find new solutions to existing problems in the field of NLP. The subject of our study is KEA itself, namely, how it can be used and what for. Thus, those applications of KEA that we will consider further represent only one of all possible varieties of approaches to solving some certain tasks, and also give new information about KEA's abilities. Despite the fact that the algorithm is not precisely new, we have chosen KEA for our experiments because it proved to be a useful and universal tool in different fields, but so far has not been used for processing Russian texts.

We would also like to state in advance that, as a significant part of the research was conducted manually, in many aspects it is not large-scale.

The paper is organized as follows. Section 2 briefly describes the structure and working principles of KEA. Section 3 contains description of the first possible KEA application, namely identification of translation equivalents, while Section 4 deals with the second experiment which concerns composing extracts for short stories based on keyphrases extracted by KEA. Section 5 is devoted to general conclusions and future work.

2 KEA Structure

KEA was developed by I.H. Witten et al. in New Zealand in 1999 [5, 6]. It is a keyphrase extraction algorithm which contains two stages:

- training: KEA is trained on the documents where the keyphrases are manually assigned by the author; as a result, a model for identifying keyphrases in new documents is created;

- extraction: the model created on the previous step is applied, and the keyphrases for new documents are identified.

On both stages, by certain rules, KEA chooses candidate phrases. The procedure of candidate selection is as follows:

1) preprocessing of the input documents:
   - tokenization;
   - relative phrase boundaries are placed;
   - non-alphabetical characters are removed.
2) keyphrase candidates filtering:
   - the length of a candidate keyphrase is limited to a certain size;
   - proper names cannot be chosen as candidate keyphrases;
   - constructions beginning or ending with a stopword cannot be candidate keyphrases.
3) case-folding and stemming.

After that for each candidate two features – *TFxIDF* and *first occurrence* – are calculated. *TFxIDF* shows how often a phrase occurs in the document in comparison to its frequency in some large corpus:

$$TF \times IDF = \frac{freq(P,D)}{size(D)} \times -log_2 \frac{df(P)}{N}, \text{where}$$

freq(*P,D*) is the number of times *P* occurs in *D*;

size(*D*) is the number of words in *D*;

df(*P*) is the number of documents of some collection of documents or in some corpus containing *P*;

N is the size of the collection or corpus.

The second feature, *first occurrence,* is the distance between a phrase first appearance and the beginning of the document, divided by the number of words in the document. The result is a number between 0 and 1.

After being trained, KEA marks each candidate as a keyphrase or non-keyphrase, which is a class future used later by Naïve Bayes classifier. Then, by applying the model built on the training stage, KEA selects keyphrases from a new document and after some post-processing operations represents the best keyphrases to a user.

When the classifier processes a candidate phrase with feature values *t* (*TF*×*IDF*) and *d* (*distance*), two quantities are calculated:

$$P[yes] = \frac{Y}{Y+N} P_{\text{TF×IDF}}[t|yes] P_{distance}[d|yes]$$

and the same for *P*[*no*], where *Y* is the number of positive instances in the train-

ing set, i.e. keyphrases assigned by the author, and *N* is the number of negative instances, i.e. candidate phrases which are not keyphrases.

The overall probability that a candidate phrase is a keyphrase, in its turn, is calculated in the following way:

$$p = P[yes]/(P[yes] + P[no])$$

According to this value, candidate keypharses are ranked and the first *r*, where *r* is a requested number of keyphrases, presented to the user.

3 Translation equivalents among Russian and English keyphrases automatically extracted by KEA

3.1 Collecting and preprocessing text corpora

Besides KEA's possible practical usages this experiment was also aimed at verifying, to which extend KEA is a language independent tool. For us it would mean that it is capable to identify conventionally 'the same words' for the same document written in several languages. For this purpose we developed a corpus using the United Nations (the UN) documents [7] as official papers have at most precise translation and are written in formal style.

The corpus contains official letters, declarations, protocols, reports, etc. On the whole, it includes 60 documents (~ 115000 tokens), where 30 documents are written in English and 30 – in Russian. In each subcorpora 25 documents were taken for the training set, while the rest 5 formed the test set. The documents in each set were picked randomly. Obviously, in the UN documents no manually assigned keyphrases are provided, so we used document-headline pairs in the training set.

As it was already mentioned, KEA is a universal language-independent algorithm that means that the importance of a phrase for the document content does not depend on any particularities of a language. Although the realization of KEA allows to provide external language-dependent modules such as stemmers, for example. And its initial package contains stemmers for some languages, but Russian is not among them. As using different stemmers for document preprocessing could influence the resulting list of keyphrases, no linguistic processing of the documents was used in this experiment. Thus, equal conditions were set up for both languages.

In processing English texts we used an internal list of stopwords, created by the developers of the algorithm, and stopword list for the Russian language was collected from Russian National Corpus (RNC) lists of function words and ab-

breviations [8]. It includes the most frequent prepositions, particles, pronouns, interjections, some parentheses, digits and Latin characters.

For each document of the test set we obtained a list of 20 (the number recommended by the developers as containing the most salient keyphrases) the most relevant keyphrases. After that the lists were manually analyzed in order to find translation equivalents.

3.2 Results and evaluation

It is worth mentioning that results obtained in the course of experiments cannot be evaluated with high precision as the algorithms of keypharse extraction as such are hard to evaluate, especially when no manually assigned keyphrases are provided. Moreover, the algorithms like KEA, as a rule, work better for the documents that were preprocessed, – for languages with rich grammar like Russian in particular. As it was already noted, we did not perform preprocessing of the documents in our study to create at most equal conditions for both languages. Therefore, for each document we decided to calculate the percentage of semantically associated structural units for both outputs combined together. The number of units being members of some kind of semantic relations was dived by 40 (20 Russian keyphrases for a document and 20 English keyphrases for a document) and multiplied by 100 to get a percentage. Technically, of course, those are two different documents, but as our study is of semantic nature, we consider it to be unimportant detail. Obtained results with examples are shown in the Table 1.

Table 1. The percentage of semantically associated structural units of a document among Russian and English keyphrases.

Document id	The percentage of semantically associated structural units of a document	Examples
G1812398\400	65%	Paris Agreement – Парижского соглашения Annex I to the Convention – приложение I к Конвенции included in Annex – включенных в приложение
G1813678\80	67,5	TIR carnet holder – держателя книжки МДП subcontractor – субподрядчика container – контейнера

N1813436\38	65%	Advisory Committee – Консультативный комитет liquidation – ликвидации mission – миссии
N1813943\46	75%	terrorism and transnational organized crime – терроризмом и транснациональной организованной преступностью Security Council – Совет Безопасности crime – преступностью
V1802422\24	60%	member states – государство-член voting rights – права голоса contributions – взносов

As we can see, we indeed can find translation equivalents in the output what proves KEA's language-independence and new possibilities for research in that area.

Although for these figures some notes should be made. Firstly, KEA tends to break semantically associated units. For instance, for the document G1812398\400 we had *Paris*, *agreement* and *Paris Agreement* for both languages. It is quite a common issue for automatic keyphrase extraction, but among researchers there is still no convention how to conduct any kind of calculations in this case. In our paper we decided to count full phrases as well as their parts. So, in the example above, all three units were considered to be semantically associated.

Secondly, because of the certain nature of texts in our corpus, we mainly dealt with translation equivalents, and sometimes it is hard to tell, whether or not keyphrases are equivalent and whether the parts came from the same phrase. For example, for a document N1813943\46 were extracted *Совет Безопасности*, *Совет Безопасности напоминает*, *Security Council* and *encourages*. In such cases we had to turn to the original text, which is not very convenient within the experiment, because it was done manually for each document in the corpus, to look at the context. But it is still impossible to tell, if *Security Council* came from *Security Council encourages* or *Security Council recalls*. As a used corpus was not aligned, looking at the context becomes a separate problem.

Therefore, such, sometimes, high figures are a product of evaluation issues appearing while processing broken phrases. Those breaks may be caused not only by KEA's peculiarities, but also by the absence of morphological preprocessing of the texts. It is commonly known that 'messy' data causes calculation mistakes, that is why we admit that our evaluation is raw and does not claim to be the only one possible or highly precise.

4 Automatic summarization of short stories

4.1 Data preprocessing

In this paper we used KEA to create extracts based on the original text [9, 10]. According to [11] extract is a collection of passages (ranging from single words to whole paragraphs) extracted from the input text(s) and produced verbatim as the summary.

For this experiment we compiled a corpus of 35 short stories written in Russian and Russian translations of famous literary works. Among the authors whose stories were used are A. Chekhov, O. Henry, D. Kharms and others. While selecting the only criterion was a small size. 30 short stories were used for the training set and the rest five for the test set. As manually assigned keyphrases for training we took abstracts for those stories written by users of [12].

Further actions can be divided in two ways:

1. Experiments based on the lemmatized training set:
   - lemmatization of the abstracts;
   - deleting stopwords;
   - lemmatization of the training set;
   - lemmatization of the test set;
   - extraction of 20 the most relevant keyphrases.
2. Experiments based on the non-lemmatized training set:
   - lemmatization of the test set;
   - extraction of 20 the most relevant keyphrases.
   - lemmatization of the output keyphrases.

The reason for this division is the fact that KEA produces different results depending on if the training set has been lemmatized or not. For lemmatization we used morphological analyzer pymorphy2 [13] in Python.

4.2 The algorithm

As the corpus has been processed and keyphrases for the test set extracted, an extract for a story is automatically composed based on obtained results. We developed and tested the algorithm which was implemented in Python. Our algorithm is composed of several modules including preprocessing as well as the module creating an extract.

The algorithm contains several stages:

1) the text is split into sentences: as the search of keyphrases in the text is conducted by lemmas, later we need to find and extract original sentences;

2) the title and the first sentence are extracted: we need the title to bound an extract with its story, and the first sentence gives it a start;
3) the search of the keyphrases in the sentences: at this point we have lemmatized original texts and their keyphrases to conduct a search by lemmas;
4) candidate sentences are assigned some scores (this stage will be discussed later);
5) selected sentences are extracted from the original text and the first five (including the first one) having a score more or equal to 2 form the extract.

Scores are assigned as follows:

1, if a keyphrase is included in one of the constructions listed below, and if it is a subject or a predicate of the sentence in the first two cases:

- noun:
  - noun + noun\verb\full adjective\short adjective (in the distance of +\- 1 from the main word)
- verb:
  - verb + noun\infinitive (in the distance of +\- 1 from the main word)
  - verb + full adjective + noun
- adjective:
  - adjective + noun (in the distance of + 1 from the main word)
  - verb + adjective + noun;

2, if a keyphrase in the sentence is among the first five from the output list;

3, if a sentence contains more than one keyphrase;

4, *5*, *6* are assigned for combinations and if a sentence contains several keyphrases.

4.3 Results and evaluation

Therefore, the obtained extracts are as follows.

Here is an extract for ‘Enemies’ by A. Chekhov. The story begins when to the doctor, whose son has just died, comes a visitor and asks for help because his wife is sick. The doctor refuses saying that he cannot work now, but eventually agrees to come.

Table 2. Examples of extracts automatically composed by the proposed algorithm.

Lemmatized training set	Non-lemmatized training set
ВРАГИ. *В десятом часу темного сентябрьского вечера у земского доктора Кирилова скончался от дифтерита его единственный сын, шестилетний Андрей.*	*ВРАГИ.* *В десятом часу темного сентябрьского вечера у земского доктора Кирилова скончался от дифтерита его единственный сын, шестилетний Андрей.* *Кирилов, как был, без сюртука, в расстегнутой жилетке, не вытирая мокрого лица и рук, обожженных карболкой, пошел сам отворять дверь.*
— Я дома, — ответил Кирилов. *— Пока ехал к вам, исстрадался душой... Одевайтесь и едемте, ради бога... Произошло это таким образом.* *— Верьте, я сумею оценить ваше великодушие, — бормотал Абогин, подсаживая доктора в коляску.* *В его осанке, в плотно застегнутом сюртуке, в гриве и в лице чувствовалось что-то благородное, львиное; ходил он, держа прямо голову и выпятив вперед грудь, говорил приятным баритоном, и в манерах, с какими он снимал свое кашне или поправлял волосы на голове, сквозило тонкое, почти женское изящество.*	*— Я дома, — ответил Кирилов.* *Очень рад, что застал... Бога ради, не откажите поехать сейчас со мной... У меня опасно заболела жена... И экипаж со мной... По голосу и движениям вошедшего заметно было, что он находился в сильно возбужденном состоянии.* *Когда Абогин еще раз упомянул про Папчинского и про отца своей жены и еще раз начал искать в потемках руку, доктор встряхнул головой и сказал, апатично растягивая каждое слово: — Извините, я не могу ехать... Минут пять назад у меня... умер сын... — Неужели?*

In this case, the second extract seems to be more appropriate, as it is more coherent and does not contain redundant information.

Now we can see a counter-example. The story is ‘Tobin’s Palm’ by O. Henry. Two friends are going to Coney Island to cut loose because one of them, Tobin,

has just been deceived and robbed by his girlfriend. There they meet a gipsy who warns Tobin to stay away from certain people and says that he will meet a person who will bring him luck. So, the rest of the story Tobin and his friend are trying to find that person.

Table 2. Examples of extracts automatically composed by the proposed algorithm (continue).

Lemmatized training set	Non-lemmatized training set
Линии судьбы.	*Линии судьбы.*
Мы с Тобином как-то надумали прокатиться на Кони-Айленд.	*Мы с Тобином как-то надумали прокатиться на Кони-Айленд.*
Промеж нас завелось четыре доллара, ну а Тобину требовалось развлечься.	*Промеж нас завелось четыре доллара, ну а Тобину требовалось развлечься.*
Кэти Махорнер, его милая из Слайго,[70] как сквозь землю провалилась с того самого дня три месяца тому назад, когда укатила в Америку с двумя сотнями долларов собственных сбережений и еще с сотней, вырученной за продажу наследственных владений Тобина — отличного домишки в Бох Шоннаух и поросенка.	*Кэти Махорнер, его милая из Слайго,[70] как сквозь землю провалилась с того самого дня три месяца тому назад, когда укатила в Америку с двумя сотнями долларов собственных сбережений и еще с сотней, вырученной за продажу наследственных владений Тобина — отличного домишки в Бох Шоннаух и поросенка.*
— Я вижу дальше, — говорит гадалка, — что у тебя много забот и неприятностей от той, которую ты не можешь забыть.	*Ну и вот мы, я да Тобин, двинули на Кони — может, подумали мы, горки, колесо да еще запах жареных зерен кукурузы малость встряхнут его.*
— Берегись, — продолжает гадалка, — брюнета и блондинки, они втянут тебя в неприятности.	*Тобин выдает ей десять центов и сует свою руку, которая приходится прямой родней копыту ломовой коняги.*

Here, the first extract is likely to be more successfully made because it gives the story a start, while from the second one it is hard to understand what happened with characters after they had arrived at Coney Island.

To give estimation to obtained results, we asked 6 experts to evaluate the texts from the following three perspectives:

- which one of two extract variations is better: lemmatized or non-lemmatized; the one better is assigned 1 score, while the other gets 0 (further was evaluated the one that got 1 at this step);
- meaningfulness: if it is impossible to get something about a story from the extract, the score for this parameter equals 0; if a reader could get at least something, 1; and if an extract is for the most part clear, 2;
- preview: whether or not a given extract can be used as a preview for a short story.

The average evaluations for each parameter are shown in Table 3. As the first parameter is a matter of preference and refers to another issue (data preprocessing), total score was calculated only for 'Meaningfulness' and 'Preview' parameters, 3 consequently being the highest point..

Table 3. Expert evaluation of obtain results.

Title	lemmatized version	non-lemmatized version	Meaningfulness	Preview	Total score
Enemies, A. Chekhov	0,2	0,8	1,5	0,8	2,3
Strictly Business, O. Henry	1	0	1,5	0,8	2,3
Tobin's Palm, O. Henry	0,8	0,2	0,8	0,7	1,5
The Man in the Case, A. Chekhov	0,2	0,8	1,3	0,8	2,2
A story about a priest, M. Zoshchenko	0,2	0,8	1,2	0,8	2

Clearly, KEA can be used as an in-between tool for composing extracts for short stories, as it has shown competitive results, gaining the average total score more than or equal to 1,5 out of 3.

Interestingly, experts, as a rule, preferred a version based on non-lemmatized data. In a way it confirms our suggestion that stemming from the source package would be better for data preprocessing.

5 Conclusions

In this paper we tried to find and test some further applications of KEA, namely identifying translation equivalents in the same text written in several languages and summarizing short stories. As we can see, KEA has managed to find the equivalents in texts and summarize stories up to its preview. That means that KEA is capable to serve as a universal and effective tool for different tasks and may be useful not only for researchers but for naive users as well.

Acknowledgements

The reported study is supported by Russian Fund of Basic Research (RFBR) grants 16-06-00529 «Development of a linguistic toolkit for semantic analysis of Russian text corpora by statistical techniques».

References

1. Kaur, J., Gupta, V.: Effective approaches for extraction of keywords. In: International Journal of Computer Science Issues, vol. 7, № 6, pp. 144–148 (2010).
2. Beliga, S.: Keyword extraction a review of methods and approaches. URL: http://langnet.uniri.hr/papers/beliga/Beliga_KeywordExtraction_a_review_of_methods_and_approaches.pdf (2014).
3. Sokolova, E. V., Mitrofanova, O. A.: Automatic Keyphrase Extraction by applying KEA to Russian texts. In: IMS 2017 Proceedings, St.-Petersburg (2017).
4. Nenkova, A., McKeown, K.: Automatic summarization. In: Foundations and Trends in Information Retrieval, vol. 5, № 2–3, pp. 103–233. (2011).
5. KEA Homepage, http://www.nzdl.org/Kea/index.html, last accessed 2018/05/27.
6. Witten, I.H., Paynter G.W., Frank E., Gutwin C., Nevill-Manning C.G.: KEA: Practical Automated Keyphrase Extraction. In: Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, IGI Global, pp. 129– 152 (2005).
7. The United Nations Homepage, http://www.un.org/ru/index.html, last accessed 2018/05/27.
8. RNC Homepage, http://www.ruscorpora.ru/, last accessed 2018/05/27.
9. Kazantseva, A., Szpakowicz, S.: Summarizing short stories. In: Computational Linguistics, vol. 36, № 1, pp. 71–109 (2010).
10. Luhn, H.P.: The automatic creation of literature abstracts. In: IBM Journal of research and development, vol. 2, № 2, pp. 159–165 (1958).
11. Hovy, E., Lin, C.Y.: Automated text summarization and the SUMMARIST system. In: Proceedings of a workshop on held at Baltimore, Maryland: October 13–15, 1998, Association for Computational Linguistics, pp. 197–214 (1998).
12. FantLab Homepage, https://fantlab.ru/, last accessed 2018/05/27.
13. Korobov, M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages. In: Analysis of Images, Social Networks and Texts, pp 320–332 (2015).

Studying Text Complexity in Russian Academic Corpus with Multi-level Annotation

Marina Solnyshkina[1], Valery Solovyev[2], Vladimir Ivanov[3], Andrey Danilov[4]

[1] Kazan Federal University, Kazan, Russia
mesoln@yandex.ru
[2] Kazan Federal University, Kazan, Russia
maki.solovyev@mail.ru
[3] Innopolis University, Kazan, Russia
nomemm@gmail.com
[4] Kazan Federal University, Kazan, Russia
tukai@yandex.ru

Abstract. The problem of compiling a large multi-level annotated corpus of Russian academic texts was sparked by the demand to measure complexity (difficulty) of texts assigned to certain grade levels in terms of meeting their cognitive and linguistic needs. For this purpose we produced a corpus of 20 textbooks on Social Studies and History written for Russian secondary and high school students. Measuring text complexity called for linguistic annotations at various language levels including POS-tags, dependencies, word frequencies. Three complexity formulas are compared as an example of using a corpus to study the complexity of texts.

Keywords: multi-level, annotated corpus, Russian academic texts, text complexity, POS-tags, dependencies, word frequencies.

1. Introduction

Automatic multi-level analysis of language implies utilizing a large corpus or a number of corpora which are viewed to be of great value for several research tasks [24]. In this paper we present the ongoing project carried out at Kazan Federal University (Russia) aimed at compiling and annotating a corpus of Russian academic texts.

To the best of our knowledge, no prior corpus-based research has been specifically conducted with the aim of estimating text complexity of Russian educational materials on Social studies. The specific, though sporadic, studies of Russian text readability did not go beyond using mere collections of limited texts of a specific type or genre: fiction (mostly for academic purposes) [17], legal [8], academic texts (chemistry, mathematics, economics) [26, 14, 20, 27]. Most of the research carried out in the area was based on English and other Germanic languages for native and/or non-native readers [3, 6, 10, 16, 22, 23]. The shortage of previous corpus-based research on text complexity of modern Russian academic texts provides a strong justification for pursuing the current study. Our objective is to introduce a multi-level annotated corpus of Russian academic texts with the ultimate goal of disseminating its potential in Russian discourse research.

It is the authors hope that this proliferation will contribute to detailed examination, identification and measurement of Russian text features. The paper is organized in the following way: In section Background we first give an introduction to the problem of text complexity, we also present the empirical approach to the problem applied in modern multidisciplinary studies. In section Corpus Description we provide information on the corpus collection regarding the type of the texts collected, the size of the corpora and the ultimate goal behind the corpus collection. In same Section we also provide information on preprocessing of the corpus and the multi-level process of the annotation. In Section 4 we briefly describe our experiments conducted with the compiled corpus and in the conclusion section we offer the authors' insights into the areas of possible utilization of the corpus and the perspectives of the work.

2. Background

The earliest studies on readability dating back to late 19th century were mostly aimed at developing readability formulas and utilized a limited number of quantitative features: average sentence length, average word length and word frequency [13, 4, 5]. Given the simplicity of the models and availability of the variables, the readability formulas have been the focus of harsh criticism since they appeared for the first time. Modern advances in natural language processing (NLP) allowed obtaining lexical and syntactic features of a text, as well as automatically train readability models using machine-learning techniques [23]. Text readability studies based of ngram models were successfully conducted by American researchers [9] and later on, based on syntax simplicity/complexity, discourse characteristics (narrativity, abstractness, referential and deep cohesion,

etc., extended to assessing a particular text profile and its target audience see [16]. Modern researchers of English develop NLP tools of new generation providing accurate and valid analyses on various dimensions of texts and measure complex discourse constructs using surface-level linguistic features such as text structure, vocabulary or the number of unique words in a text, givenness or the number of determiners and demonstratives in a text, anaphor or the number of all pronouns lexical diversity, connectives and conjuncts which together with anaphor are indicators of text coherence, future as an indicator for situational cohesion, syntactic complexity measured through the number of words per sentence, and the number of negations [7]. Based on systemic language parameters text features are to be specified for one language only. Thus, every modern NLP tool as well as a readability formula are applicable to one language in particular. E.g. parameters measured for English cannot be applied to estimating Russian texts complexity as Germanic languages have limited morphology in comparison with Russian [23] and all text features need to be validated in a corpus of a considerable size. Owing to the existing lack of available corpora Russian discourse studies at the moment are viewed as underdeveloped [25]. Russian academic texts began being used in readability studies only in 1970-s [21], but with a short break during 1990-s the studies in the area were quite extensive. Nowadays researchers view the following text readability features as cognitively significant: number of syllables, number of words, sentence count, average sentence length, abstract words count, homonyms counts, polysemous words counts, technical terms counts, etc. [20]. Ivanov V.V. tested correlations of 49 factors, among which the strongest correlations are identified for the percentage of short adjectives, the percentage of finite verb form, the Flesch-Kincaid Grade Level Score, the Flesch Reading Ease Score [13], the Coleman and Liau index, average number of words per sentence, percentage of complex sentences, percentage of compound sentences, percentage of abstract words [11]. Karpov N. et al. [26] conducted a series of experiments utilizing a number of machine-learning models to automatically rank Russian texts based on their complexity. For the purpose the authors compiled two subcorpora: (1) a corpus of texts generated by teachers for learners of Russian as a foreign language (at http://texts.cie.ru); (2) 50 original news articles for native readers. They assessed 25 text parameters of each text in the corpora, such as sentence length, word length, vocabulary, parts of speech classification. For the last fifteen years, readability of Russian academic texts has been actively discussed at conferences in Russia and abroad as well as in numerous publications [21] but readability studies are still far from being systematic and irregularities in reporting make it difficult to draw firm conclusions [23] mostly due to corpora limitations. The problem of defining Russian text complexity features can be studied on a massive corpus containing academic

texts used in modern schools. Unfortunately neither Russian National Corpus nor Corpora of Russian (http://web-corpora.net/?l=en) though being large and widely used in studies of lexical, syntactic and discourse features cannot be used for the purposes of our research based on the fact that they do not provide access to modern Russian academic texts.

3. Corpus Description

For the purposes of the study we compiled a corpus of two sets of textbooks on Social Studies and History written for Russian secondary and high school students. The total size of the corpus of 20 textbooks is more than 1 million tokens.

The first collection of 14 texts from textbooks on Social Studies by Bogolubov L. N. marked "BOG" by Nikitin A.F. marked "NIK" aimed for 5 – 11 Grade Levels. In our study, Grade Levels means the class number for which the textbook is intended. It was selected to teach the predictive model and define independent variables of the text variation. The second collection of 6 texts from textbooks on History by different authors aimed for 10 – 11 Grade Levels. Both sets of textbooks are from the "Federal List of Textbooks Recommended by the Ministry of Education and Science of the Russian Federation to Use in Secondary and High Schools".

To ensure reproducibility of results, we uploaded the corpus on a website thus providing its availability online. Note, however, that the published texts contain shuffled order of sentences. The sizes of BOG and NIK subcollections of texts are presented in Table 1.

Table 1. Properties of the preprocessed corpus on Social Studies

	Tokens		Sentences		Words per sentence	
Grade	BOG	NIK	BOG	NIK	BOG	NIK
5-th	--	17,221	--	1,499	--	11.49
6-th	16,467	16,475	1,273	1,197	12.94	13.76
7-th	23,069	22,924	1,671	1,675	13.81	13.69
8-th	49,796	40,053	3,181	2,889	15.65	13.86
9-th	42,305	43,404	2,584	2,792	16.37	15.55
10-th	75,182	39,183	4,468	2,468	16.83	15.88

10-th*	98,034	--	5,798	--	16.91	--
11-th	--	38,869	--	2,270	--	17.12
11-th*	100,800	--	6,004	--	16.79	--

In the Table 1 star sign (*) denotes advanced versions of books for the corresponding grade; sign '-' denotes absence of a textbook for the corresponding grade.

Data on the collection of books on history is presented in Table 2. The first column lists textbook authors and the class number.

Table 2. Properties of the preprocessed corpus on History

Author / Grade	Tokens	Sentences	Words per sentence
Soboleva / 10-th	81544	7116	11.46
Volobuyev 10-th	40949	3676	11.14
Guryanov / 11-th	100331	9393	10.68
Petrov / 11-th	85409	8536	10.01
Plenko / 11-th	63804	5292	12.06
Ponomarev / 11-th	44833	4003	11.2

3.1 Corpus Preprocessing

For the convenience, we have preprocessed all texts from the corpus in the same way. Common preprocessing included tokenization and splitting text into sentences. During the preprocessing step we excluded all extremely long sentences (longer than 120 words) as well as too short sentences (shorter than 5 words) which we consider outliers. Clearly, such sentences can be not outliers at all in another domain, but for the case of school textbooks on Social Studies sentences shorter than 5 words are outliers. Sentence and word-level properties of the preprocessed dataset are presented in Tables 1 and 2.

Extremely short sentences mostly appear as names of chapters and sections of the books or as a result of incorrect sentence splitting. We omit those sentences, because the average sentence length is a very important feature in text complexity assessment and hence should not be biased due to splitting errors. At the same time sentences with five to seven words in Russian can still be viewed as short sentences, because the average sentence length (in our corpus) is higher than ten.

Table 1 demonstrates that values of Word per sentence (ASL) as it is generally expected, increase with the grades.

3.2 Multi-level Annotations in Corpus

All annotations in the corpus are performed on three levels: text-level, sentence-level and word-level. At the text-level meta-annotations refer to a number of sentences and a set of tokens, an author and a grade-level of a given text. At the word-level we have part-of-speech tag for each word. POS-tagging has been performed with the use of the TreeTagger for Russian (http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger/). The tagset is available from the web-site of the project. As example we provide distribution of major PoS-tags among texts on Social Studies, Table 3. We also annotate each lemma in the corpus with its relative frequency measured in the large corpus of Russian texts, Russian National Corpus.

At the sentence-level the corpus contains annotations of sentence boundaries, the tokens are assigned to sentences as well as a dependency tree of each sentence. For dependency parsing we use pretrained neural models (https://github.com/MANASLU8/ CoreNLPRusModels) for Stanford Dependency Parser for Russian (https://nlp.stanford.edu/software/stanford-dependencies.shtml). Finally, at the moment, we are adding semantic annotations to the corpus. The semantic annotations are based on the very large Russian Thesaurus (RuThes) [28]. Concepts of the RuThes are mapped to the Wordnet thesaurus that allows to process textual content at semantic level.

Table 3. Unique words in each of four PoS-tags that appear in textbooks; normalized by 1000 words

	NOUN		VERBS		ADJECTIVES		ADVERBS	
Grade	BOG	NIK	BOG	NIK	BOG	NIK	BOG	NIK
5-th	--	69.7	--	48.6	--	77.6	--	10.7
6-th	69.1	69.4	48.8	42.2	81.2	96.6	11	11.3
7-th	71.4	63.6	39.5	37.8	100.3	90.8	9.3	9.9
8-th	43	53.5	22.2	27.9	111.3	114.6	6.1	7
9-th	38.3	46.5	21.3	24.2	119.4	114.8	5.5	6.6
10-th	33.5	50.1	17.3	22.8	124.5	130.6	4.4	6.6
10-th*	28.6	--	14.7	--	122.3	--	4	--
11-th	--	43.4	--	23	--	124.2	--	6.2
11-th*	30.7	--	14	--	143.7	--	3.9	--

4. Studies of Text Readability and Complexity

First of all, the corpus can be used to adjust readability formulas in Russian. Second, even very simple statistics provided in the Table 3 can be useful in text complexity studies. For example, one can see that average number of unique adjectives grow when grade level increases. At the same time average number of adverbs (as well as verbs) decreases. Both observations correspond with idea that texts become more descriptive. However, with assistance of the data it is possible to measure the correlation.

In this study, 3 formulas (our formulas [29], Matskovskiy Readability Formula [30] and Oborneva's Readability Formula [17]) were applied to 5 Social Studies and 7 History textbooks for grades 10 – 11. In the formulas below, GL denote the grade level.

In paper [29] we provided readability formula GL = 0.36ASL + 5.76ASW – 11.97, where ASL and ASW means average of words per sentence and means average of syllables per word respectively. Below, this formula is labeled RRF. In [30] Matskovskiy M.S. computed the first readability formula for the Russian language: GL = 0.62ASL + 0.123X + 0.051, where X is the percentage of three syllable words in the text. In [17] Oboroneva I. introduced readability formula readability formula GL = 0.5ASL + 8.4 ASW – 15.59.

In an attempt to verify the features defined as contributing to text readability but not measured by the existing readability formulas, we compared the 11 texts under study in order to see what metrics better correlate with the grade level. The data are presented in table 4.

The Fig. 1 below shows, that Oboroneva's formula positioned them as textbook comprehensible only by people with at least 16 – 17 years of formal schooling, i.e. with Bachelor or Master's Degree. It is clear from the table that grade level predictions based upon the equation of regression of Oborneva I. do not coincide with the actual grade levels, the difference is marked in 6 years in the case of textbooks on History. As for Matskovskiy's Readability formula which was initially developed to compute readability of media texts only, it proves to be quite reliable in assessing readability of academic texts also (compare columns 'Grade' and 'Matskovskiy' in Table 4).

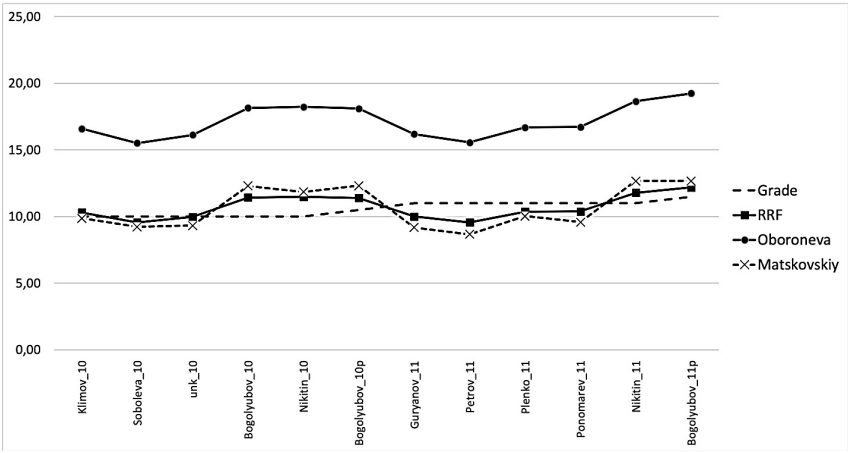


Fig. 1. Predictions of grade levels. Ground truth is represented with a dashed line

Table 4. Comparison of three readability formulas using Social Science and History textbooks

Book	ASL	ASW	Fraction of 3-sylables words	TRUE_ GRADE	RRF	Oboroneva	Matskovskiy
Guryanov_11	11.14	3.12	0.18	11.00	10.01	16.19	9.19
Klimov_10	12.45	3.09	0.17	10.00	10.31	16.60	9.88
Petrov_11	10.43	3.09	0.18	11.00	9.57	15.56	8.67
Plenko_11	12.52	3.10	0.18	11.00	10.38	16.69	10.03
Ponomarev_11	11.64	3.15	0.19	11.00	10.39	16.73	9.59
Soboleva_10	11.75	3.00	0.15	10.00	9.57	15.53	9.23
BOG_10	15.88	3.07	0.20	10.00	11.44	18.15	12.31
BOG_10*	16.06	3.06	0.19	10.50	11.41	18.11	12.33
BOG_11*	16.03	3.19	0.22	11.50	12.19	19.25	12.68
NIK_10	15.06	3.13	0.20	10.00	11.49	18.24	11.85
NIK_11	16.19	3.11	0.21	11.00	11.79	18.66	12.68

5. Discussion

Thus, there are two reasons which make future research into Russian texts readability relevant. First, the recent reports from educators call for improving read-

ing comprehension in secondary and high schools throughout the country [2, 1]. Researchers also testify to Russian students lack of interest in reading caused by inappropriate selection of educational materials [20]. The Corpus is a valuable instrument for discourse studies as its data and flexible search system provide a solid foundation for comparative research of modern Russian texts and enables deep insights into patterns and dependencies of different text features. The Corpus is also viewed by the authors as a powerful tool for discovering new aspects and regularities of Russian discourse.

Acknowledgements

This research was financially supported by the Russian Science Foundation, grant # 18-18-00436, the Russian Government Program of Competitive Growth of Kazan Federal University, and the subsidy for the state assignment in the sphere of scientific activity, grant agreement 34.5517.2017/6.7. The Russian Academic Corpus (section 3, 3.1 in the paper) was created without support from the Russian Science Foundation.

References

1. Kompetentnostnyy podkhod v vysshem professionalnom obrazovanii (pod redaktciyey A.A. Orlova, V.V. Gracheva), Tula (2012).
2. Berezhkovskaja E. Problema psihologicheskoj negotovnosti k polucheniju vysshego obrazovanija u studentov mladshih kursov. M.: Prospec. (2017).
3. Britton, B.K., & Gulgoz, S. Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. Journal of Educational Psychology, 83, pp. 329-404 (1991).
4. Chall J., Dale E. Readability revisited: The new Dale-Chall readability formula. Brookline Books (1995).
5. Coleman M., Liau T. L. A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60:283-284 (1975) .
6. Cornoldi, C., & Oakhill, J. (Eds.). Reading comprehension difficulties: Processes and intervention. Hillsdale, NJ: Erlbaum (1996).
7. Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D. S. Analyzing discourse processing using a simple natural language processing tool (SiNLP). Discourse Processes, 51, pp. 511-534 (2014).
8. Dzmitryieva A. Iskusstvo yuridicheskogo pis'ma: kolichestvennyy analiz resheniy Konstitutsionnogo Suda Rossiyskoy Federatsii [The art of legal writing: a quantitative analysis of the Russian Constitutional Court rulings]. Sravnitel'noe konstitutsionnoe obozrenie, no.3, pp. 125-133. (In Russian) (2017).

9. Heilman M., Thompson K. C., Callan J., and Eskenazi M. Combining lexical and grammatical features to improve readability measures for first and second language texts. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-07), pp. 460-467, Rochester, New York (2007).
10. Jackson, G. T., Guess, R. H., & McNamara, D. S. Assessing cognitively complex strategy use in an untrained domain. In N. A. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), Proceedings of the 31st Annual Meeting of the Cognitive Science Society. pp. 2164-2169. Amsterdam, The Netherlands: Cognitive Science Society (2009).
11. Ivanov V. K voprocu o vozmonosti ispolzovanija lingvisticeskix xarakteristik slonosti teksta pri issledovanii okulomotornoj aktivnosti pri ctenii u podrostkov [Toward using linguistic profiles of text complexity for research of oculomotor activity during reading by teenagers]. Novye issledovanija [New studies], 34(1):4250 (2013).
12. Karpov N., Baranova J., and Vitugin F. Single-sentence readability prediction in Russian. In Proceedings of Analysis of Images, Social Networks, and Texts conference (2014).
13. Kincaid J. P., Fishburne R. P. Jr., Rogers R. L., and Chissom B. S. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. Research Branch Report 8-75, Naval Technical Training Command, Millington, TN (1975).
14. Krioni N., Nikin A., and Fillipova A. Avtomatizirovannaja sistema analiza slozhnosti uchebnyh tekstov [The automated system of the analysis of educational texts complexity]. Vestnik UGATU (Ufa), 11(1):28 (2008).
15. McNamara, D.S. Reading both high and low coherence texts: Effects of text sequence and prior knowledge. Canadian Journal of Experimental Psychology, 55, pp. 51-62 (2001).
16. McNamara, D.S., Kintsch, E., Songer, N.B., & Kintsch, W. Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition & Instruction, 14, pp. 1-43 (1996).
17. Obobroneva I. Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov. M.: RAS Institut soderzhaniya i metodov obucheniya (2006).
18. Okladnikova S. Modelkompleksnoj ocenki citabelnosti testovyx materialov na etape razrabotki [A model of multidimensional evaluation of the readability of test materials at the development stage]. Prikaspijskij journal: upravlenie i vysokie texnologii, 3:6371 (2010).
19. Popova Ja.I., Shishkevich E.V. Standartizacija uchebnoj literatury srednej shkoly po kriteriju udobochitaemosti In Sevastopol'skij nacional'nyj universitet jadernoj jenergii i promyshlennosti. Nauchnye vedomosti BelGU. Ser. Gumanitarnye nauki. 12. No. 6. pp. 142-147 (2010).
20. Shpakovskiy Y. et al. Otsenka trudnosti vospriyatiya i optimizatsiya slozhnosti uchebnogo teksta. PhD thesis (2007).

21. Solnyshkina M., Harkova E, and Kiselnikov A. Comparative coh-metrix analysis of reading comprehension texts: Unified (Russian) state exam in English vs Cambridge first certificate in English. English Language Teaching, 7(12):65 (2014).
22. Ozuru, Y., Rowe, M., OReilly, T., & McNamara, D. S. Wheres the difficulty in standardized reading tests: The passage or the question? Behavior Research Methods, 40, pp. 1001-1015 (2008).
23. Reynolds R. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories, San Diego, CA: 16 June 2016. In: Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications, pp.289-300 (2016).
24. Sinclair, J. Corpus Evidence in Language Description, in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.) Teaching and Language Corpora. London/New York: Longman, pp. 27-39 (1997).
25. Ivanov V.V., Solnyshkina M.I., & Solovyev V.D. Efficiency of text readability features in Russian academic texts. In Computational Linguistics and Intellectual Technologies, V.17, pp. 277–287 (2018).
26. Karpov N., Baranova J., and Vitugin F.. Single-sentence readability prediction in Russian. In International Conference on Analysis of Images, Social Networks and Texts, pp. 91-100. Springer (2014).
27. Ustinova, L. V., Fazylova L. S. Avtomatizacija ocenki slozhnosti uchebnyh tekstov na osnove statisticheskih parametrov. Vestnik Karagand. un-ta. Ser. Matematika. 1. pp. 96-103 (2014).
28. Loukachevitch, N. V., Lashevich, G., Gerasimova, A. A., Ivanov, V. V., & Dobrov, B. V. Creating Russian Wordnet by conversion. Kompjuternaja Lingvistika i Intellektualnye Tehnologii, 15, pp. 405-415 (2016).
29. Solovyev V., Ivanov V., & Solnyshkina M. Assessment of reading difficulty levels in Russian academic texts: Approaches and Metrics, Journal of Intelligent & Fuzzy Systems, vol.34, no.5, pp. 3049-3058 (2018).
30. Matskovskiy, M.S. Problemy chitabelnosti pechatnogo materiala [Problemsof printed material readability]. In: Dridze, T.M. & Leontev, A.A. (eds) Smyslovoe vospriyatie rechevogo soobshcheniya v usloviyakh massovoy kommunikatsii [Semantic perception of verbal communication in the context of mass communication]. Moscow: Nauka (1976).

An overview of the available corpora for evaluation of the automatic keyword extraction algorithms

A. Vanyushkin[1][0000-0002-2556-6305] and L. Graschenko[2][0000-0002-7972-1358]

[1] Pskov State University, Pskov 180000, Russia
alexmandr@mail.ru
[2] Institute of Mathematics named A. Juraev of the Academy of Sciences of the Republic of Tajikistan, Dushanbe 734063, Republic of Tajikistan
graschenko@mail.ru

Abstract. The article discusses the evaluation of automatic keyword extraction algorithms (AKWEA) and points out AKWEA's dependence on the properties of the test collection for effectiveness. As a result, it is difficult to compare different algorithms which tests were based on various test datasets. It is also difficult to predict the effectiveness of different systems for solving real-world problems of natural language processing (NLP). We considered six publicly available analytical text collections, since analytical articles are typical for the keyword extraction task. Our analysis revealed that their text length distributions are very regular and described by the lognormal form. Moreover, most of the article lengths range between 400 and 2500 words. Then we take in to consideration a number of characteristics, such as the text length distribution in words and the keyword assignment method, of eleven corpora. All these corpora are significantly different from each other in such characteristics as their text length distribution, size, themes and authorship of the keyword assignment, but were used in keyword extraction evaluation tasks. Only one of them, DUC-2001, has the most relevant form and distribution parameters but its disadvantage is the small number of experts participating in the keyword assignment. Moreover, all the corpora are monolingual and do not allow carry cross-language study.

Keywords: Text Corpus, Corpus Linguistics, Keyword Extraction, Text Length Distribution, Natural Language Processing, Information Retrieval.

1 Introduction

The number of digital documents available is growing on a daily basis at an over-whelming rate. As a consequence, there is a need to increase the complexity of the structure and software solutions in the field of NLP which are based on a number of basic methods and algorithms. The algorithms of automatic keyword and key phrase (KW) extraction are among them. This task has been analyzed over the past sixty years from different perspectives. There has been a significant increase in the number of researches that took place in the last twenty years, of which many have been publications of different AKWEA's [30]. The reason for this is the increasing amount of computing research, data resources and especially the development of internet services. It also simplifies the development and evaluation of new algorithms.

The term "keyword" is interdisciplinary and above all, is used in works on psycho-linguistics and Information Retrieval [32] that causes the existence of different approaches to its definition. Summarizing the numerous opinions, we can conclude that the keywords (phrases) are words (phrases) in the text that are especially important, commonly understood, capacious and representative of a particular culture. The set of which can give a high-level description of its content for the reader and providing a compact representation and storage of its meaning in mind [30]. In practice, the terms keyword and key phrase have the same meaning.

Despite the large amount of specialized and interdisciplinary work there has not been a consistent technique developed for detecting keywords yet. Experiments confirmed that this is done intuitively by people, and is personality, and even gender-based [20]. This implies the non-triviality of the development of formal methods and KW extraction algorithms for computing. Therefore, the current efforts of researchers are focused on the development and implementation of hybrid learning-based AKWEA's which assumes the use a variety of linguistic resources. Thus, the accuracy of training and control datasets has great importance on the effectiveness of development.

Our analysis reveals number problematic areas. The author's results in testing AKWEA's are often different from those obtained by other researchers, since they use different control data in the evaluation of algorithms [30]. Independent testing of KW extraction algorithms is a difficult task because there is a lack of implemented system and source code of algorithms in open access. This problem is partially solved by carrying out workshops when the organizers propose test data for all participants. At the same time the number of available and well-proven corpora for KW extraction evaluation is small (10-20) and the criteria for their formation are not methodologically well enough investigated. The possibility

of transferring the results of the algorithms in other languages remains an open question. The remarkable thing is that most of the known results are obtained for the English language, and the rules for the interpretation of them to the Slavic languages, especially to Russian, have not been established.

Indeed, preliminary empirical data show that for the graph-based algorithms with increased text size the precision of AKWEA's might reduce. Therefore, the effective-ness of the algorithms depends on the type and parameters of the text lengths distribution (in words) that constitute research data. Homogeneity of the data by genre and text difficulty probably has some influence on the effectiveness of AKWEA's too (see Fig. 1).

A separate discussion is necessary to explore the characteristics of experimental corpora such as size, existence and the methods of KW assignment (who and how many authors assigned them), the subject and the type of text (abstracts and full articles). KW assignment can be performed by authors, experts on the topic or by crowdsourcing. In this case, questions arise such as what kind of assignment is considered optimal, is it possible to rely on public opinion and what is a minimum number of participants that must specify the word as a keyword to assign it as such. It should be noted that the quality of KW assignment depends on the size of a corpus. As the size increases, the complexity of assignment rises.

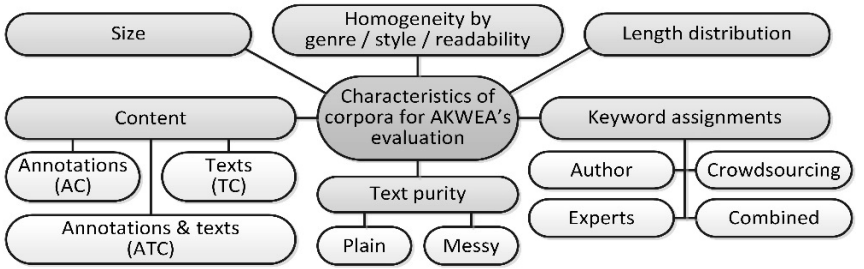


Fig. 1. The specifications of research corpora for keyword extraction evaluation.

But first of all, it is necessary to investigate existing text collections (those used for KW extraction) for the length distribution parameters (in words).

2 Methodology and Research Tools

Articles from six web sites were selected as the statistical and research database subset that contains a voluminous collection on various English topics. This

choice is due to the assumption that the main sphere of work for KW extraction is mostly with topical or subject-based text, especially those that contain elements of analytical themes. The eleven corpora (test and trial), that were used in some or other research or scholarly articles, were found using a search engine.
Many sites block automatic downloading for article collection or don't have freely available archives for use at all. So, sites with freely available resources were used. After downloading the collection of articles, automatically parsing of the pages was made and the text was extracted. Then the tokenization and a count of the number of words in each article was made. Stanford Log-linear Part-Of-Speech Tagger[1] was used for tokenization of English texts, which is widely used in both research and commercial sectors [12].

The text lengths distributions in words were presented for every collection. We used Pearson's chi-squared test to evaluate the fitness of observed data to some theoretical distributions using advanced analytics software package Statistica[2] and EasyFit[3] software. It is worth pointing out that the form distribution depends on the mode of data grouping [11]. Calculating the number of bins k in different ways leads to a wide range of its possible values. For the expected Gaussian distribution, the Sturges formula is normally used, but if the data are not normal or there are more than 200 cases, it's poorly applied [7].

For the unification of the calculation the bin sizes in the histograms we used the *Freedman and Diaconis* rule, which gives the value agreed with the recommendations on standardization[4] and then convert it into the number of bins:

$$h = 2(IQ)n^{-1/3} \quad (1)$$

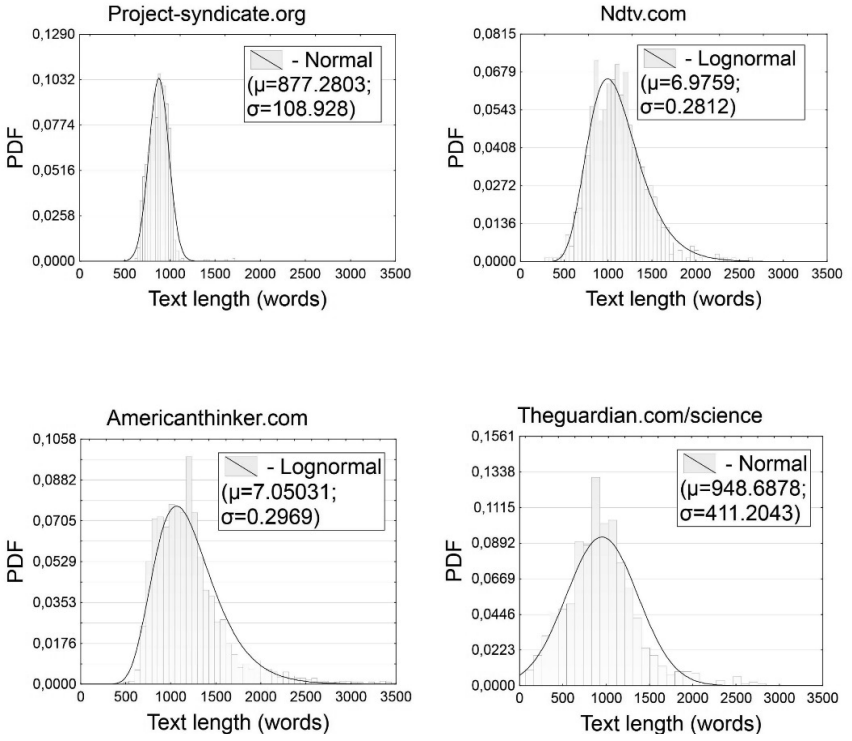
where h is the bin size, IQ is the interquartile range of the data and n is the number of observations. At the same time according to the Pearson's chi-squared test (p-value = 0.05) we did not obtain a satisfactory fit of the results in all cases. Our hypothesis was confirmed by varying k in a small range with respect to the calculated value. To improve the accuracy of estimates of the form and parameters of the probability density function further research is needed. For example, the Levenberg-Marquardt algorithm was used by other researches to solve similar problems [27].

[1] http://nlp.stanford.edu/
[2] http://www.statsoft.com
[3] http://www.mathwave.com/
[4] R 50.1.033-2001. Applied statistics. Rules of check of experimental and theoretical distribution of the consent. Part I. Goodness-of-fit tests of a type chi-square

3 A Review of Existing Information Resources

3.1 Text Length Distributions in Analytical Articles Collections

The issue of natural length distribution and optimal lengths are taken into consideration by many researches. Most studies have been devoted to investigate blog post sizes [8, 21, 29], which describes the text length distribution with fat tails. This is true for the user comments [27], e-mail messages [22] and for the length of the texts that are stored on users' computers [3]. It is proposed [2] to consider the length of the articles from Wikipedia encyclopedia as an indicator of their quality, and the overall length of the English papers described by the lognormal form [26]. Fig. 2 presents the probability density function distributions for the six data-sets.



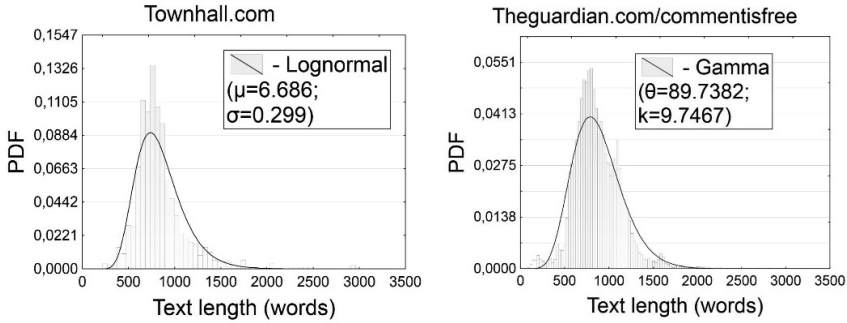


Fig. 2. Distribution of analytical articles lengths in words

As can be seen from the graphs, the majority of the length distribution of analytical articles can be comparative to the normal or lognormal form. The majority of texts are in the range of 400 to 2500 words.

Table 1 presents general information and statistical characteristics of the reviewed text collections. Collection size ranges from 736 to 14529 articles and their publication dates cover the period from 2015 to 2016. Mean lengths of articles varies between 839-1212 words.

Table 1. Characteristics of the analytical article collections

№	Source	Count	Text length				Publishing period
			Mean	Min.	Max.	Std. Dev.	
1	project-syndicate.org	1163	873,3	612	1721	108,9	01.15-12.15
2	ndtv.com	736	1112,5	274	2650	309,9	01.15-12.15
3	americanthinker.com	2268	1212,2	473	3703	410,4	01.15-02.15
4	townhall.com	905	839,5	217	2960	283,9	07.15-12.16
5	theguardian.com/science	897	948,7	66	2848	411,2	01.15-12.16
6	theguardian.com/ commentisfree	14529	874,6	79	3045	278,8	01.15-12.16

It is worth pointing out that there are possible restrictions authors can have on the length of published articles. For example, on project-syndicate.org a recommended article length by their editorial team is 1000 words.

3.2 Existing Corpora for Keyword Extraction Evaluation

Despite the large number of works devoted to keyword extraction evaluation the number of specially trained and public corpora are much less so. Some of them are used multiple times in different studies. *Hulth-2003* [6] for example, consisting of abstracts of scientific articles, is one of the most popular and was used in the many academic papers [5, 18, 23-25, 28, 33]. Other datasets are used much less frequently, often only by their authors. One of the main drawbacks of such corpora is the "messy" texts, as many of them contain a bibliography, tables, captions and pictures in text files.

We surveyed eleven public corpora, which are significantly different from each other such as the text length distribution as well as other characteristics such as the size, themes and authorship of the keyword assignment. Table 2 summarizes the characteristics of reviewed corpora. The following are some explanations.

Table 2. Characteristics of the available corpora for KW extraction evaluation

№	Corpus	Year	Contents	KW assign	Type	Resource
1	DUC-2001 [31]	2001	News articles	E-2	ATC	github.com
2	Hulth-2003 [6]	2003	Paper abstracts from Inspec	E-?	AC	researchgate.net
3	NLM-500 [1, 4]	2005	Full papers of PubMed documents	E-?	ATC	github.com
4	NUS [19]	2007	Scientific conference papers	A+E-?	ATC	github.com
5	WIKI-20 [15, 16]	2008	Technical research reports of computer science	E-15	ATC	github.com
6	FAO-30 [14, 16]	2008	Documents from UN FAO[1]	E-6	TC	github.com
7	FAO-780 [14, 16]	2008	Documents from UN FAO	E-?	TC	github.com
8	KRAPIVIN [10]	2009	ACM[2] full papers 2003-2005	A	ATC	disi.unitn.it

[1] Food and Agriculture Organization
[2] Association for Computing Machinery

9	CiteULike [16, 17]	2009	Bioinformatics papers	O-3	TC	github.com
10	SemEval-2010 [9]	2010	ACM full papers	A+E-0,2	ATC	github.com
11	500N-KPCrowd-v1.1[13]	2012	News articles	O-20	TC	github.com

Note: notation of KW assignment: A-text authors, O-N – Crowdsourcing (N – number of people per one text, ? - n/a), E-experts.

Let us explain the features of the KW assignment of the given corpora. *DUC-2001* was prepared for text summarization evaluation within the Document Understanding Conferences, but KW assignment was made by two only graduate students in 2008 for the study of AKWEA's [31]. A feature of the *Hulth-2003* assignment is the presence of two sets of KW – a set of controlled, i.e. terms restricted to the Inspec thesaurus, and a set of uncontrolled terms that can be any terms. *NLM-500* sets of keywords restricted to the thesaurus of Medical Subject Headings. *WIKI-20* assigned by 15 teams consisting of two senior computer science undergraduates each. These KW sets were restricted to the names of Wikipedia articles. *NUS* has the author's assigned KW lists as well as KW lists assigned by student volunteers.

FAO-30 and *FAO-780* differ in size and composition of the experts, but both KW sets were restricted to the Agrovoc[1] thesaurus. In *KRAPIVIN* parts of the articles are separated by special characters, which makes it convenient to their separate processing. *CiteULike* KW's were assigned by 322 volunteers but the authors noted that for this reason the high quality of the KW assignment is not guaranteed. For assignment of 500N-KeyPhrasesCrowdAnnotated-Corpus (*500N-KPCrowd-v1.1*) the researchers used the crowdsourcing platform Amazon's Mechanical Turk[2].

SemEval-2010 has been specially prepared for the Workshop on Semantic Evaluation 2010, where 19 systems were evaluated by matching their KW's against manually assigned ones. It consists of three parts: trial, training and test data. The authors note that on average 15% of the reader-assigned KW and 19% of the author-assigned KW's did not appear in the papers.

Table 3 shows the statistical characteristics of text length distributions in the reviewed corpora.

[1] http://www.fao.org/agrovoc
[2] https://www.mturk.com/

Table 3. Statistical characteristics for the datasets used in this paper.

№	Name	Count	Mean	Min.	Max.	Std. Dev.
1	DUC-2001	307	769,1	141	2505	435,1
2	Hulth-2003	2000	125,9	15	510	59,9
3	NUS	211	6731,7	1379	13145	2370,6
4	NLM-500	500	4805	436	24316	2943,3
5	WIKI-20	20	5487,8	2768	15127	2773,4
6	FAO-30	30	19714,3	3326	70982	16101,6
7	FAO-780	779	30106,5	1224	255966	31076,5
8	KRAPIVIN	2304	7572,8	144	15197	2092,3
9	CiteULike	180	6454,1	878	23516	3408,9
10	SemEval-2010	244	7669,1	988	13573	2061,9
11	500N-KPCrowd-v1.1	447	425,9	38	1478	311,7

Fig. 3–7 shows the text length distributions of the reviewed corpora.

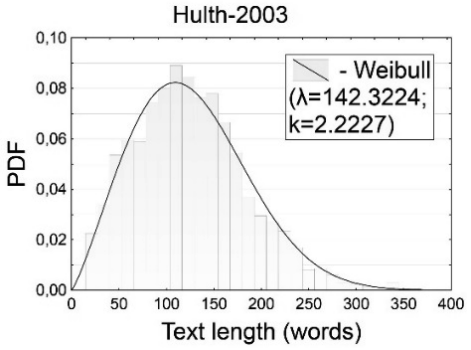


Fig. 3. Distribution of annotation lengths in words

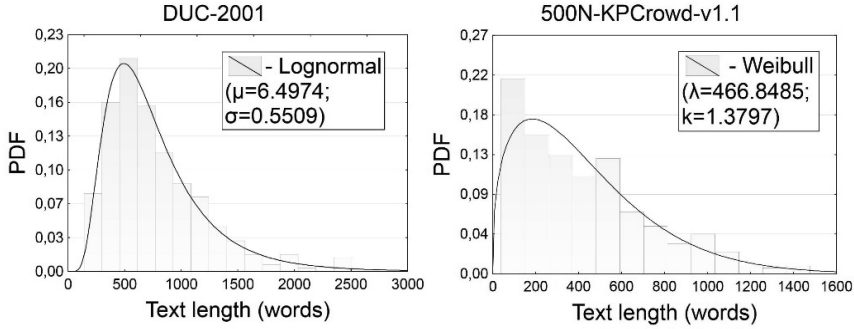


Fig. 4. Distribution of news article lengths in words

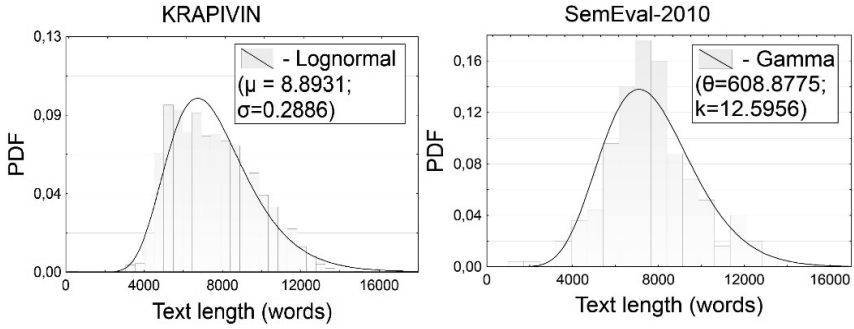


Fig. 5. Distribution of ACM article lengths in word

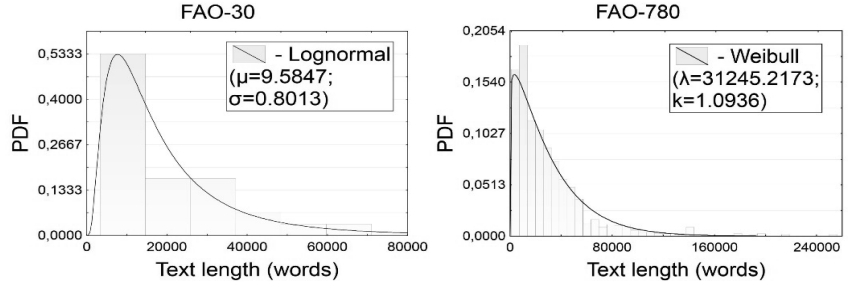


Fig. 6. Distribution of FAO document lengths in words

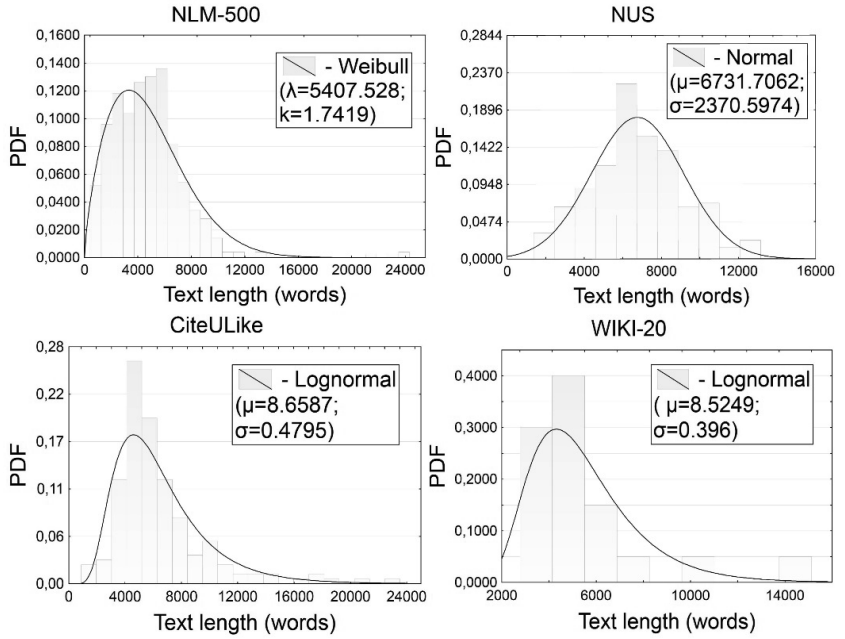


Fig. 7. Distribution of Scientific paper lengths in words

A review of test corpora revealed that they differ significantly on the sizes, the themes, and the method of keyword assignment. The difference of text lengths for some couples is three orders of magnitude. The text length in the tens of thousands of words questioned the possibility and the meaning of the use of AKWEA's at its entire length, without division into semantic parts. In contrast, annotation in definition contain a higher percentage of KW's than text containing a few thousand words.

The text length distribution histograms of the most reviewed corpora have outliers, and does not correspond to the established in Section 3.1 principles, that is their apparent drawback. *DUC-2001* has the most relevant form and distribution parameters (LN (6.49, 0.55)) but its disadvantage is the small number of experts participating in the KW assignment (only two). Moreover, all the above corpora are monolingual and do not allow carry cross-language study of KW extraction.

4 Conclusions

As can be seen from the above, the majority of the texts for which KW extraction is relevant are in the range of 400 to 2500 words and their text length distribution is quite well described by the lognormal form. Thus, in practice it is advisable to use AKWEA's that show a good performance in certain text length ranges. However, in general a comparison of existing AKWEA's was performed on corpora with different characteristics. Moreover, the length of the manually assigned KW lists in them varies widely, and KW assignment was made by different categories of people such as students, volunteers and experts for example. Thus, for an objective comparison of existing AKWEA, it is necessary to use corpora, whose characteristics are close to those of natural collections.

References

1. Aronson, A.R., Mork, J. G., Clifford, W.G., Humphrey, S.M., Rogets, W.J.": The NLM Indexing Initiative's Medical Text Indexer. Studies in Health Technology and Informatics 107, 268–272 (2004).
2. Blumenstock, J.E.: Size matters: word count as a measure of quality on wikipedia. Proceedings of the 17th International Conference on World Wide Web, pp. 1095–1096, Beijing (2008).
3. Douceur, J.R., Bolosky, W.J.: A Large-Scale Study of File-System Contents. Proceedings of the 1999 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, pp. 59–70, Atlanta (1999).
4. Gay, C.W., Kayaalp, M., Aronson, A.R.: Semi-automatic indexing of full text biomedical articles. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, pp. 271–275 (2005).
5. Hasan, K.S., Ng, V.: Conundrums in unsupervised keyphrase extraction: making sense of the State-of-the-Art. Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference, pp. 365–373, Beijing (2010).
6. Hulth, A.: Improved Automatic Keyword Extraction Given More Linguistic Knowledge. Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing, pp. 216–223, Sapporo (2003).
7. Hyndman, R.J.: The problem with Sturges' rule for constructing histograms,http: http://robjhyndman.com/papers/sturges.pdf, last accessed 2018/07/01 (1995).
8. Kagan, N.: Why Content Goes Viral: What Analyzing 100 Million Articles Taught Us, http://okdork.com/why-content-goes-viral-what-analyzing-100-millions-articles-taught-us, last accessed 2018/04/05 (2014).
9. Kim, S., Medelyan, O., Kan, M., Baldwin, T.: Semeval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 21–26, Los Angeles (2010).

10. Krapivin, M., Autayeu, A., Marchese, M.: Large Dataset for Keyphrases Extraction, http://eprints.biblio.unitn.it/archive/00001671/01/disi09055-krapivin-autayeu-marchese.pdf, last accessed 2018/07/01 (2009).
11. Lemeshko, B.Y., Postovalov, S.N.: Limit distributions of the Pearson $\chi 2$ and likelihood ratio statistics and their dependence on the mode of data grouping. Industrial laboratory 64(5), 344–351 (1998).
12. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics, pp. 55–60, Baltimore (2014).
13. Marujo, L., Gershman, A., Carbonell, J.G., Frederking, R.E, Neto, J.P.: Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing. In 8th International Conference on Language Resources and Evaluation (LREC 2012), pp. 399–403, Istanbul (2012).
14. Medelyan, O., Witten, I.H.: Domain Independent Automatic Keyphrase Indexing with Small Training Sets. Journal of the American Society for Information Science and Technology 59(7), 1026–1040 (2008).
15. Medelyan, O., Witten, I.H., Milne, D: Topic Indexing with Wikipedia. Proceedings of the Wikipedia and AI workshop at AAAI-08, pp. 19–24, Chicago (2008).
16. Medelyan, O.: Human-competitive automatic topic indexing (Thesis), The University of Waikato, Hamilton (2009).
17. Medelyan, O., Frank, E., Witten, I.H.: Human-competitive tagging using automatic keyphrase extraction. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 1318–1327, Singapore (2009).
18. Mihalcea, R., Tarau, P.: TextRank: Bringing order into texts. Proceedings of EMNLP 2004, pp. 404–411, Barcelona (2004).
19. Nguyen, T., Kan, M.: Keyphrase extraction in scientific publications. Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers, pp. 317–326, Hanoi (2007).
20. Nozdrina, T.G.: Reconstructing original textes by keywords [Osobennosti vosstanovlenija tekstov – originalov na osnove kljuchevyh slov], Modern problems of science and education [Sovremennye problemy nauki i obrazovanija] 1-2, 167-174 (2015).
21. Patel, N, Why 3000+ Word Blog Posts Get More Traffic (A Data Driven Answer), http://neilpatel.com/blog/why-you-need-to-create-evergreen-long-form-content-and-how-to-produce-it/, last accessed 2018/05/22.
22. Paxson, V.: Empirically-Derived Analytic Models of Wide-Area TCP Connections. IEEE/ACM Transactions on Networking 2(4), 316–336 (1994).
23. Popova, S.V., Khodyrev, I.A.: Tag lines extraction and ranking in text annotation [Izvlechenie i ranzhirovanie kljuchevyh fraz v zadache annotirovanija]. Scientific and technical journal of information technologies, mechanics and optics [Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki] 1, 81-85 (2013).
24. Rousseau, F., Vazirgiannis, M.: Main core retention on graph-of-words for single-document keyword extraction. ECIR, pp. 382–393, Vienna (2015).

25. Schluter, N.: Centrality Measures for Non-Contextual Graph-Based Unsupervised Single Document Keyword Extraction. In Proceedings of TALN 2014, pp. 455–460, Marseilles (2014).
26. Serrano, M.A., Flammini, A., Menczer, F.: Modeling statistical properties of written text. PLoS One 4(4), 1–8 (2009).
27. Sobkowicz, P., Thelwall, M., Buckley, K., Paltoglou, G., Sobkowicz, A.: Lognormal distributions of user post lengths in Internet discussions - a consequence of the Weber-Fechner law?. EPJ Data Science 2, 1–20 (2013).
28. Tsatsaronis, G., Varlamis, I., Nørvåg, K.: SemanticRank: Ranking Keywords and Sentences Using Semantic Graphs. Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1074–1082, Beijing (2010).
29. Tunguz, T.: The Optimal Blog Post Length to Maximize Viewership, http://tomtunguz.com/content-marketing-optimization, last accessed 2018/05/25 (2013).
30. Vanyushkin, A.S., Graschenko, L.A.: Methods and algorithms of keyword extraction [Metody i algoritmy izvlecheniya klyushevyh slov]. New information technology in the automated systems [Novye informacionnye tekhnologii v avtomatizirovannyh sistemah], pp. 85–93, Moscow (2016).
31. Wan, X., Xiao, J.: Single Document Keyphrase Extraction Using Neighborhood Knowledge. Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, pp. 855–860, Chicago (2008).
32. Yagunova, E.V.: Experiment and calculation in the analysis of literary text's keywords [Eksperiment i vychisleniya v analize klyuchevyh slov hudozhestvennogo teksta]. Collection of scientific works of the department of foreign languages and philosophy of PSC of UB RAS. Philosophy of Language. Linguistics. Linguodidactics [Sbornik nauchnyh trudov kafedry inostrannyh yazykov i filosofii PNC UrO RAN. Filosofiya yazyka. Lingvistika. Lingvodidaktika], 85-91 (2010).
33. Zesch, T., Gurevych, I.: Approximate Matching for Evaluating Keyphrase Extraction. Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing, pp. 484–489, Borovets (2009).

Improving the Quality of Information Retrieval Using Syntactic Analysis of Search Query

Nadezhda Yarushkina[1][0000-0002-5718-8732], Aleksey Filippov[1][0000-0003-0008-5035] and Maria Grigoricheva[1][0000-0001-7492-5178]

[1] Ulyanovsk State Technical University,
Ulyanovsk, Sev. Venec str., 32, 432027, Russia,
jng@ulstu.ru, al.filippov@ulstu.ru, gms4295@mail.ru

Abstract. The paper describes the methods for linguistic analysis of search queries to improve the quality of information retrieval. The description of the parse tree in the form of a structure containing information about two or more related words with the indication of their parts of speech and location in the original request is used to the translation of search query in Elasticsearch Query DSL. Elasticsearch Query DSL has several disadvantages: the user may not know the features of Elasticsearch Query DSL, words joined by the OR operator is using by default in information retrieval. The using of the OR operator unnecessarily increases the recall and reduces the precision of information retrieval. Taking into account the features of Elasticsearch Query DSL and the information needs of the user allow to improve the quality of information retrieval.

Keywords: Information Retrieval, Syntactic Analysis, Search Queries.

1 Introduction

Information retrieval is the process of searching in an extensive collection of data some semi-structured (unstructured) material (document) that satisfies the information needs of the user.

Semi-structured data is data that does not have a clear, semantically noticeable and easily distinguishable structure. Semi-structured data is the opposite of structured data. The canonical example of structured data is relational databases. Relational databases are typically used by enterprises to store product registers, employee personal data, etc [1,2,3].

The quality of search in information retrieval systems is usually characterized by two criteria: recall and precision. The total count of found documents determines the recall. The ratio between the determines the precision found relevant documents and the total count of documents [2,3].

The quality of the search is affected by both the characteristics of the information retrieval system itself and the quality of the search query. An ideal search query can be formed by a user who knows well the domain area. Also, to form an ideal query, the user needs to know the features of current information retrieval system and their information retrieval query language. Otherwise, the search result will have the low precision or low recall values [1,2,3].

2 Main problem

To information retrieval, the user formulates a search query. The search query is a formalized way of expressing the information needs of information retrieval system users. Information retrieval query language is used for the expression of information needs. The syntax of information retrieval query language varies from system to system. Modern information retrieval systems allow entering a query in natural language in addition to an information retrieval query language [1]. Information retrieval system finds documents containing the specified keywords or words that are in any way related to the keywords based on the user search query. The result of the information retrieval system is a list of documents, sorted by relevance [2,3].

In this paper will consider the work of the proposed method on the example of the information retrieval system Elasticsearch [4].

Elasticsearch provides a full Query DSL [5]. The query string is parsed into a series of terms and operators. A term can be a single word – *quick* or *brown* – or a phrase, surrounded by double quotes – *"quick brown"* – which searches for all the words in the phrase, in the same order.

Operators allow to customize the search:

1. By default, all terms are optional, as long as one term matches. A search for *foo bar baz* will find any document that contains one or more of *foo* or *bar* or *baz*.
   The preferred operators are + (this term must be present) and - (this term must not be present). All other terms are optional. For example, this query:
   quick brown +fox –news
   states that:

- *fox* must be present;

- *news* must not be present;
- *quick* and *brown* are optional – their presence increases the relevance.

2. Multiple terms or clauses can be grouped together with parentheses, to form sub-queries:
   (quick OR brown) AND fox.

Therefore, the Elasticsearch search algorithm has several disadvantages:

1. The user may not know the features of Elasticsearch Query DSL.
2. Words joined by the OR operator are used by default in information retrieval. The using of the OR operator unnecessarily increases the recall of information retrieval and reduces its precision.

It is necessary to develop a method of linguistic analysis and translation of a search query into a search query in a format of Elasticsearch Query DSL. The new format of search query allows to take into account the features of Elasticsearch Query DSL and improve the quality indicators (precision and recall) of information retrieval.

3 The method of Linguistic Analysis of Search Query for Improving Quality of Information Retrieval

The primary goal of the developed method of linguistic analysis and translation of a search query into a search query in a format of Elasticsearch Query DSL is the improvement of information retrieval quality. The main task is to select in the search query the groups of terms, united by some semantics.

3.1 The method of linguistic analysis and translation of a search query

The scheme of linguistic analysis of texts does not depend on the natural language itself. Regardless of the language of the source text, its analysis goes through the same stages [6,7]:

1. Splitting the text into separate sentences.
3. Splitting the text into separate words.
4. Morphological analysis.
5. Syntactic analysis.
6. Semantic analysis.

The first two stages are the same for most natural languages. Language-specific differences usually appear in the processing of word abbreviations, and in the processing of punctuation marks to determine the end of a sentence.

The results of the syntactic analysis are used to select in the search query the groups of terms, united by some semantics. To identify a noun phrase from a query identification of important query terms is necessary. It is also necessary to define the relationship between the terms of the query. A noun phrase or nominal phrase is a phrase that has a noun (or indefinite proper noun) as its head or performs the same grammatical function as such a phrase. Noun phrases are ubiquitous cross-linguistically, and they may be the most frequently occurring phrase type.

SyntaxNet as an implementation of the syntactic analysis process is used. SyntaxNet is a TensorFlow-based syntax definition framework that uses a neural network. Currently, 40 languages including Russian are supported. The source code of the already-trained Parsey McParseface neural network model that is suitable for parsing text is published For TensorFlow. The main task of SyntaxNet is to make computer systems able to read and understand human language. The precision of the model trained in the SinTagRus case is estimated at 87.44% for the LAS metric (Label Attachment Score), 91.68% for the UAS metric (Unlabeled Attachment Score) and determines the part of speech and the grammatical characteristics of words with an accuracy of 98.27% [8,9,10,11,12,13].

It is necessary to parse the search query to obtain a parse tree on the first step of the algorithm. To obtain data about the search query structure, dependencies between words and the types of these dependencies the resulting parse tree will be used.

The parse tree can be represented as the following set:

$$T = \{t_1,\ t_2,\ \ldots,\ t_k\}, \tag{1}$$

where k is a count of nodes in the parse tree;

t_i is a node of the parse tree, can be described as:

$$t_i = \left(i,\ w_i,\ m_j,\ c\right),\ i = \overline{1,k},$$

where i is an index of the word in search query;

$w_i \in W$, $W\{w_1,\ w_2,\ \ldots,\ w_k\}$, W is a set of words of search query;

$m_j \in M, M = \{Noun,\ \mathrm{Pr}opernoun,\ Verb,\ Adverb,\ Adjective,\ Conjunction,\ \mathrm{Pr}eposition,\ Interjection\}$ is a set of parts of speech for natural language;

c is an index of the word in the search query, that depends to the i -th word.

Thus, the search query is converted into a parse tree on the first step of the algorithm. For each word in the search query the part of speech, index of this word in the search query, and relations with other words of the search query are set.

The description of the parse tree in the form of a structure containing information about two or more related words with the indication of their parts of speech and location in the original request is used on the second step of the algorithm.

In the process of analysis of the input parse tree, the nodes that reflect the semantics of this query are selected. Search and translation of selected nodes into Elasticsearch Query DSL is executed using a set of rules. The translation process uses a set of rules. Rules are used to add special characters from the Elasticsearch Query DSL to the words of the search query. Also, stop words are deleted from search query during the translation. The result of the algorithm is a new search query that takes into account the semantics of information needs and features of the Elasticsearch Query DSL.

This algorithm can be represented as the following equation:

$$F^{Query} : (T, R) \rightarrow Q^*,$$

The input parameters of the function F^{Query} are the parse tree of search query T (eq. 1) and the set of rules R, and the result is a translated query Q^*.

$R = \{R_1, R_2, \ldots, R_n\}$ is the set of rules for searching elements in parse tree and their translation in Elasticsearch Query DSL.

Each rule can be represented as the following expression:

$$R_i(p, t_1, t_2, \ldots, t_m) = Q_j^*,$$

where p is a rule priority. The priority of the rule determines the order in which the rules are applied to the search query. Thus, the priority allows applying the more "complex" rules to define complex phrases first. If the more "complex" rule did not work, the rule with a lower priority is applied;

t_k is k-th element of the rule that allows to selecting the node (nodes) of the parse tree to process;

m is a count of elements in the rule;

$Q_j^* \in Q^*, j = \overline{1, q}$ is an element of translated query. Each element of a translated query contains the word or words of the original search query escaped by a symbol from the set of Elasticsearch Query DSL operators.

3.2 Examples of rules for linguistic analysis and translation of a search query

The formal description of the rule to search the noun phrase in the search query can be represented as follows:

$$R_{noun\_phrase} = \big(4,\ \langle i,\ w_i,\ Noun, c\rangle,\ \langle i+1,\ w_{i+1},\ Adjective,\ i\rangle,$$
$$\langle i+2,\ w_{i+2},\ Adjective,\ i\rangle,\ \ldots,\ \langle i+d,\ w_{i+2},\ Adjective,\ i\rangle\big) =$$
$$= +"w_{i+1}\ w_{i+2}\ \ldots\ w_{i+d}\ w_i",$$

where d is a count of adjectives in the noun phrase.

Extraction of noun phrase from the parse tree of the search query finds one or more adjective that subordinates to the current noun tree node. The result of this rule is a noun phrase, escaped with ‘ +” ‘ at the beginning and ‘ “ ‘ at the end.

Each rule consists of the two sides: the left side and the right side. The left side is a template of a parse tree fragment. The right side is a string template. The method extracts fragments of the parse tree, matched by the pattern in the left side of the rule, and then converts them to the strings escaped by the symbols of Elasticsearch Query DSL, filling the string template.

The formal description of the rule to search the related nouns in the search query can be represented as follows:

$$R_{noun\_noun} = \big(3,\ \langle i,\ w_i,\ Noun, c\rangle,\ \langle i+1,\ w_{i+1},\ Noun,\ i\rangle,$$
$$\langle i+2,\ w_{i+2},\ Noun,\ i\rangle,\ \ldots,\ \langle i+d,\ w_{i+2},\ Noun,\ i\rangle\big) =$$
$$= +"w_{i+1}\ w_{i+2}\ \ldots\ w_{i+d}\ w_i",$$

where d is a count of nouns that related to i -th noun.

Extraction of related nouns from the parse tree of the search query finds one or more noun that subordinates to the current noun tree node. The result of this rule is a set of nouns, escaped with ‘ +” ‘ at the beginning and ‘ “ ‘ at the end.

The formal description of the rule to search the proper noun that subordinates to the noun in the search query can be represented as follows:

$$R_{noun\_proper\_noun} = \big(2,\ \langle i,\ w_i,\ Noun, c\rangle,\ \langle i+1,\ w_{i+1},\ \mathrm{Pr}oper\ noun,\ i\rangle,$$
$$\langle i+2,\ w_{i+2},\ \mathrm{Pr}oper\ noun,\ i\rangle,\ \ldots,\ \langle i+d,\ w_{i+2},\ \mathrm{Pr}oper\ noun,\ i\rangle\big) =$$
$$= +"w_i\ w_{i+1}\ w_{i+2}\ \ldots\ w_{i+d}',$$

where d is a count of proper nouns that related to i -th noun.

Extraction of proper nouns that subordinates to the noun from the parse tree of the search query find one or more proper noun that subordinates to the current noun tree node. The result of this rule is a set of nouns, escaped with ‘ +” ‘ at the beginning and ‘ “ ’ at the end.

The formal description of the rule to search the proper noun in the search query can be represented as follows:

$$R_{proper\_noun} = \big(1,\ \langle i,\ w_i,\ \mathrm{Pr}\,oper\ noun, c\rangle,\ \langle i+1,\ w_{i+1},\ \mathrm{Pr}\,oper\ noun,\ i\rangle,$$
$$\langle i+2,\ w_{i+2},\ \mathrm{Pr}\,oper\ noun,\ i\rangle,\ \ldots,\ \langle i+d,\ w_{i+2},\ \mathrm{Pr}\,oper\ noun,\ i\rangle\big) =$$
$$= +"w_{i+1}\ w_{i+2}\ \ldots\ w_{i+d}\ w_i',$$

where d is a count of proper nouns that related to i -th proper noun.

Extraction of proper nouns from the parse tree of the search query finds one or more proper noun that subordinates to the current proper noun tree node. The result of this rule is a set of nouns, escaped with ‘ +” ‘ at the beginning and ‘ “ ‘ at the end.

The formal description of the rule to search the single noun in the search query can be represented as follows:

$$R_{single\_noun} = \big(1,\ \langle i,\ w_i,\ Noun,\ c\rangle\big) = +w_i.$$

As the main problem of the search subsystem of the SOM, users called the low quality of the information retrieval. This rule has a lower priority and is executed after the rules with a higher priority have been skipped only. The rule allows finding a node of parse tree with a part of speech noun that is not associated with other nodes with a part of speech noun, adjective, or proper noun.

The formal description of the rule to search the verb in the search query can be represented as follows:

$$R_{verb} = \big(1,\ \langle i,\ w_i,\ Verb,\ c\rangle\big) = +w_i.$$

This rule has the highest priority and does not overlap with any other rule. This rule allows finding a node of the parse tree with a part of speech verb.

4 Experiments

To test the method of linguistic analysis of search query proposed in this study some experiments were conducted. Figure 1 shows the primary form of application for translation of search query in Elasticsearch Query DSL.

The control “Original Query” is used to input original search query. The parse tree for original search query will be shown in table “Search query formatting elements” after pressing the button “Prepare query”. Each line of this table contains the name of the triggered rule and potential semantic group. Panel “Translation rules” allows the user to select necessary rules for translation the

original search query to query in a format of Elasticsearch DSL. The resulting search query will be shown in control "Search query in new format" after pressing the button "Format".

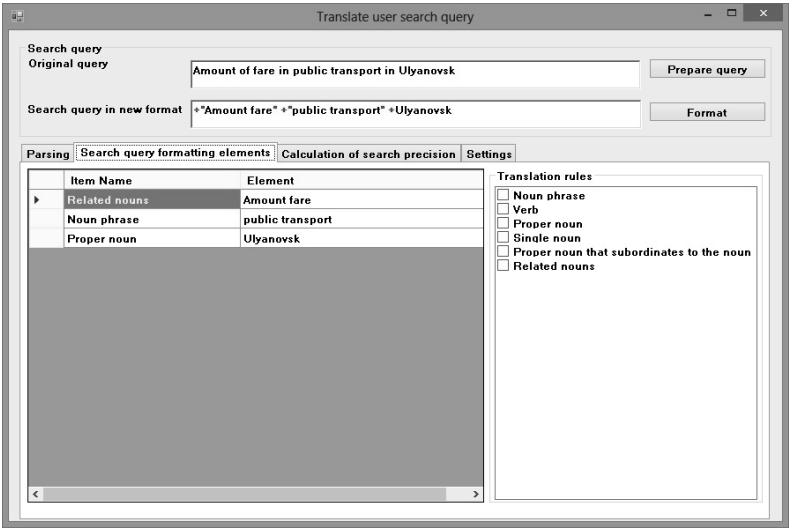


Fig. 1. The main form of application for translation of search query in Elasticsearch Query DSL

In this paper will consider the work of the proposed method on the example of the existing information retrieval subsystem of the system for opinion mining in social media (SOM). The SOM consists of the following subsystems:

1. Subsystem for importing data from external sources. This subsystem works mass media sites. Mass media loader retrieves data from HTML pages of mass media sites based on rules. The creation of own rules for each mass media is needed. The rule should contain a set of CSS-selectors. The ontology loader allows loading ontologies in OWL or RDF format into the data storage subsystem. Ontology is used for a description of the features of the problem area.
2. The data storage subsystem provides the representation of information extracted from mass media in a unified structure that is convenient for further processing. The data is stored in the context data sources, versions, etc. As database management systems are used:

- Elasticsearch for indexing and retrieving data;

- MongoDB for storing data in JSON format;
- Neo4j for storing graphs of social interaction (social graph) and ontology.

3. The data converter converts the data imported from mass media into an internal SOM unified structures.
4. The OWL/RDF-ontology translator translates ontology into the graph representation [14].
5. The semantic analysis subsystem performs preprocessing of text resources. Also, this subsystem performs statistical and linguistic analysis of text resources.
6. The information retrieval subsystem finds objects related to a specific search query. In this case, the search query can be semantically extended using an ontology.

Posts and comments downloaded from mass media sites *ulpressa.ru*, *ulgrad.ru* and *mosaica.ru/ru/ul* are used as the dataset. The collection of documents is in Russian. The proposed rules are designed for the Russian language.

As the main problem of the search subsystem of SOM, users called the low quality of information retrieval.

Thus, the precision indicator of information retrieval is used to assess the quality of the proposed method. The recall and F-measure values are not used because the data storage subsystem of SOM contains the large count of documents.

Figure 2 shows the parse tree for search query: *"Стоимость проезда на общественном транспорте в Ульяновске"*. For example, will use the query *"Amount of fare in public transport in Ulyanovsk"* in English, which is close in meaning and structure to the query in Russian. The nodes of the parse tree are the words of the search query. Each node is assigned a part of speech.

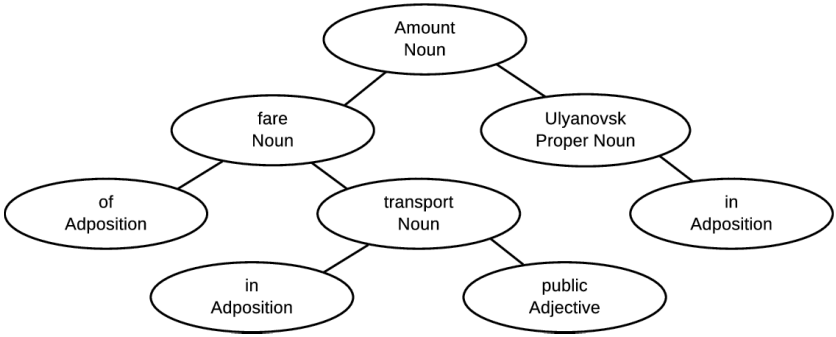


Fig. 2. The parse tree for the search query "Amount of fare in public transport in Ulyanovsk"

After work of the algorithm, the significant elements were found in the parse tree (fig. 3). In the resulting tree (fig. 3), the nodes labeled as a rule with that they were found.

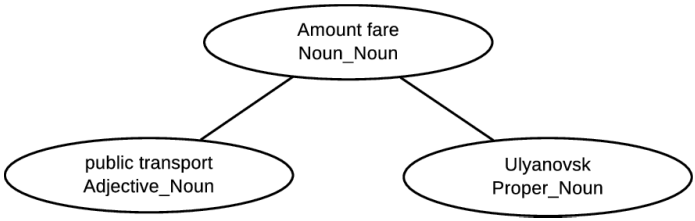


Fig. 3. The result tree for the search query *"Amount of fare in public transport in Ulyanovsk"*

Thus, after linguistic analysis and translation the resulting search query for search query *"Amount of fare in public transport in Ulyanovsk"* is *'+"Amount fare" +"public transport" +Ulyanovsk'*.

The precision value is calculated using the following expression:

$$P = \frac{a}{b}, \tag{2}$$

where a is a count of relevant documents in the search result;

b is a total count of documents in the search result.

For search query Q^O *"Amount of fare in public transport in Ulyanovsk"* the count of relevant documents in the search result of the information retrieval is 8. The total count of documents is 44857. Thus, the precision $P\left(Q^O\right)$ of the information retrieval for search query *"Amount of fare in public transport in Ulyanovsk"* is (eq. 2):

$$P\left(Q^O\right) = \frac{8}{44857} = 0,00018.$$

For search query Q^T *'+"Amount fare" +"public transport" +Ulyanovsk'* translated from search query *"Amount of fare in public transport in Ulyanovsk"* using the proposed method the count of relevant documents in the search result of the information retrieval is 8. The total count of documents is 8. Thus, the precision $P\left(Q^T\right)$ of the information retrieval for search query *'+"Amount fare" +"public transport" +Ulyanovsk'* is (eq. 2):

$$P\left(Q^T\right) = \frac{8}{8} = 1.$$

Thus, the using of proposed method improve the precision of information retrieval because of reducing the count of documents in the search result.

5 Conclusion

Elasticsearch Query DSL has several disadvantages:

1. The user may not know the features of Elasticsearch Query DSL.
2. Words joined by the OR operator are used by default in information retrieval. The using of the OR operator unnecessarily increases the recall and reduces the precision of information retrieval.

A method of linguistic analysis and translation of search query in Elasticsearch Query DSL allows improving the precision of information retrieval.

The search query is converted into a parse tree on the first step of the algorithm. For each word in the search query the part of speech, index of this word in the search query, and relations with other words of the search query are set. The description of the parse tree in the form of a structure containing information about two or more related words with the indication of their parts of speech and location in the original request is used on the second step of the algorithm. The result of the algorithm is a new search query that takes into account the semantics of information needs and features of the Elasticsearch Query DSL.

According to the results of 20 computational experiments, we can conclude: the use of the proposed method allows to increase the precision of information retrieval by an average of 18 times. The proposed algorithm can be used with any information retrieval system because it preprocessed the original search query, but does not change the system parameters or logic of information retrieval. However, the method requires adaptation to the features of the information retrieval query language of the current information retrieval system.

Acknowledgments

The study was supported by:

- the Ministry of Education and Science of the Russian Federation in the framework of the project No. 2.1182.2017/4.6. Development of methods and means for automation of production and technologicalpreparation of aggregate-assembly aircraft production in the conditions of a multi-product production program;
- the Russian Foundation for Basic Research (Grants No. 18-47-730035 and 16-47-732054).

References

1. Voorhees, E. M.: Natural language processing and information retrieval. In: Information Extraction, pp. 32-48. Springer, Berlin, Heidelberg (1999)
2. Manning C., Raghavan P., Schütze H.: Introduction to Information Retrieval. Cambridge University Press. (2008)
3. Miwa M. The Past, Present and Future of Information Retrieval: Toward a User-Centered Orientation. http://www.archives.go.jp/english/news/pdf/151106miwa_en.pdf. Accessed 20 Oct 2018
4. Elasticsearch. https://www.elastic.co/. Accessed 20 Oct 2018
5. Elasticsearch Query DSL. https://www.elastic.co/guide/en/elasticsearch/reference/2.3/query-dsl-query-string-query.html. Accessed 20 Oct 2018
6. SRILM - The SRI Language Modeling Toolkit. http://www.speech.sri.com/projects/srilm. Accessed 20 Oct 2018
7. Manning C., Schutze H.: Foundations of Statistical Language processing. In: The MIT Press. (1999)
8. Sboev A.G., Gudovskikh D.V., Ivanov I., Moloshnikov I.A., Rybka R.B., Voronina I.: Research of a Deep Learning Neural Network Effectiveness for a Morphological Parser of Russian Language. http://www.dialog-21.ru/media/3944/sboevagetal.pdf, Accessed 20 Oct 2018 (2017)
9. Chang J., Seefried J., Taylor S., Brandner A. SyntaxNet: Google's Open-sourced Syntactic Parser. http://www.sfs.uni-tuebingen.de/~keberle/NLPTools/presentations/SyntaxNet/ SyntaxNet.pdf. Accessed 20 Oct 2018
10. Petrov S. Announcing SyntaxNet: The World's Most Accurate Parser Goes Open Source. https://ai.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html. Accessed 20 Oct 2018
11. Weiss D., Petrov S. An Upgrade to SyntaxNet, New Models and a Parsing Competition. https://ai.googleblog.com/2017/03/an-upgrade-to-syntaxnet-new-models-and.html. Accessed 20 Oct 2018
12. Alberti C., Orr D., Petrov S. Meet Parsey's Cousins: Syntax for 40 languages, plus new SyntaxNet capabilities. https://ai.googleblog.com/2016/08/meet-parseys-cousins-syntax-for-40.html. Accessed 20 Oct 2018
13. Eberle K. Natural Language Tools - Test and Comparison. http://www.sfs.uni-tuebingen.de/~keberle/NLPTools/slides/slides01.pdf. Accessed 20 Oct 2018
14. Yarushkina N., Filippov A., Moshkin V.: Development of the Unified Technological Platform for Constructing the Domain Knowledge Base Through the Context Analysis. In: Creativity in Intelligent Technologies and Data Science, pp 62-72 (2017)

Pragmatic Markers in the Corpus "One Day of Speech": Approaches to the Annotation

K.D. Zaides[1], T.I. Popova[2], N.V. Bogdanova-Beglarian[3]

[1-3] Saint Petersburg State University, Saint Petersburg, Russia
kristina.zaides@student.spbu.ru, tipopova13@gmail.com,
n.bogdanova@spbu.ru

Abstract. The article describes the scheme of the annotation of pragmatic markers in the corpus of Russian everyday speech "One Day of Speech". Pragmatic markers are defined as special units in the speech that have only pragmatic function without any (or with 'bleached') lexical meaning. The annotation of pragmatic markers is usually performed manually due to the existing ambiguity of markers in different contexts. The typology of pragmatic markers includes different groups marked with special annotation tags. The annotation process was split into two stages since several issues of tagging of PMs arose. The main problems, which occurred during the annotation process, and the possible ways of their solution are also discussed in the research. The paper propose the improved methods of problem solving during the annotation of pragmatic markers applied to the corpus of oral speech, which can be useful for the linguistic annotation of any other levels of oral speech.

Keywords: Pragmatic Marker, Spoken Speech, Corpus of Everyday Speech, Corpus Linguistics, Corpus Annotation.

1 Introduction

The annotation of any corpus is the main linguistic tool in the corpus structure used for receiving correct search results and meta-information about texts and authors (speakers). Nowadays, the number of corpora of oral speech is growing exponentially around the world, so that an important and relevant issue in modern linguistics is being stated–to develop the basic principles of speech annotation, including such its units, which have never been described in the scien-

tific literature before. Besides the well-known widespread levels of annotation, such as the marking of prosodic units, the part-of-speech tagging, the syntactic and semantic parsing, certain linguistic information should be tagged for some modern research tasks in communication studies, in particular, the discourse and pragmatic annotations. While the automatic annotation of a corpus material is implemented by the number of special parsers, the pragmatic annotation is still carried out manually because the instruments for such annotation are awaited to be produced in the near future [1, 2]. Moreover, many kinds of pragmatic annotation involves such patterns and details of speech that cannot be fulfilled by the automatic device, e.g., speech acts analysis or pragmatic markers revealing. This paper presents the results of two stages of pragmatic markers annotation; therefore, we focus on the definition of the term *pragmatic marker* and its characteristics below.

A *pragmatic marker* (PM) is a relatively new term in the linguistics, introduced in this meaning by N.V. Bogdanova-Beglarian [3], which is used towards the particular speech units: words, expressions and phrases fulfilling different pragmatic functions in the discourse. The meaning of a term *discourse marker* (DM) do not coincide with the content of the term *pragmatic marker* since they describe different groups of discourse/pragmatic units, although both of them demonstrate the ability to structure the discourse but by different means. Discourse markers usually either navigates the paragraphs of a text or reveal time, causal, conditional and numerous other relations between the fragments being meaningful content words with a certain lexical meaning. A brief literature review, based on different researchers' understanding of DMs, can identify the specificity of these units more narrowly.

B. Fraser defines the DM as "a pragmatic class, lexical expressions drawn from the syntactic classes of conjunctions, adverbials, and prepositional phrases" [4]. The representatives of this class mainly "signal a relationship between the segment they introduce, S2, and the prior segment, S1" [Ibid.]. Basically, according to B. Fraser, they fall into two types: "those that relate aspects of the explicit message conveyed by S2 with aspects of a message, direct or indirect, associated with S1; and those that relate the topic of S2 to that of S1" [Ibid.]. The researcher characterizes the DM as "a linguistic expression only which: (i) has a core meaning which can be enriched by the context; and (ii) signals the relationship that the speaker intends between the utterance the DM introduces and the foregoing utterance" [Ibid.]. As it is explained, "they function like a two-place relation, one argument lying in the segment they introduce, the other lying in the prior discourse" [Ibid.]. Syntactically, DMs do not form a separate syntactic category. So-called *pragmatic markers* B. Fraser earlier identified as "structures and expressions which linguistically encode aspects of the speaker's

direct communicative intention" [5] that "do not contribute to the propositional content of the sentence but signal different types of messages" [4].

D. Schiffrin argues that DMs do not fit completely into some linguistic category since their main function lies in adding to discourse coherence and providing "contextual coordinates for ongoing talk" [6]: DMs are "sequentially dependent elements which bracket units of talk" [Ibid.] which can be sentences, prepositions, speech acts, tone units, etc.

L. Schourup describes as DMs "conversational particles such as *well* and *oh*, parenthetical lexicalized clauses such as *y'know* and *I mean*, and a variety of connective elements in speech and writing, including *so*, *after all*, and *moreover* [7]. L. Schourup pointed out that "DMs are more often regarded as comprising a functional class that draws on items belonging to various syntactic classes" [Ibid.].

E. Traugott notices that DMs "allow speakers to display their evaluation not of the content of what is said, but of the way it is put together, in other words, they do metatextual work". [8]. The author supposes that DMs (in this work, the markers *indeed*, *in fact*, *besides* are investigated) go the grammaticalization path from the clauseinternal adverbial through the sentence adverbial to the discourse particle, the subtype of the class of discourse markers [Ibid.].

In case of the annotation, the hesitation disfluencies sometimes are classified as discourse markers [9]. We suppose that such approach is not very productive since the hesitations can be detected automatically and usually treated as phonetically filled hesitation pauses and not as markers.

To the contrast, pragmatic markers derive from both content and functional words (nouns, verbs, adverbs, prepositions, etc.), and, during the process not only of grammaticalization, but also of pragmaticalization, they lose (in whole or in part) their lexical and/or grammatical meaning and get pragmatic one in some of their everyday speech usages. A content or functional word becomes a PM in a process of pragmaticalization: as a result, the role of its pragmatic component increases and a role of significant component decreases. The pragmatic function of a PM turns to be the leading one for a certain word, wherein the grammatical component can be still presented (for example, Aijmer reports that some units like *I think* are pragmaticalized, but they still have tense, aspect, and mood [10]). In this understanding, pragmatic markers such as *you know*, *I think*, *sort of*, *actually*, and *that sort of thing*, "have the function of checking that the participants are on the same wavelength or of creating a space for planning what to say making revisions, etc." [Ibid.]. PMs in the discourse approach "express speaker attitude to what has gone before, what follows, the discourse situation, and so forth" [8]. The further development of a pragmatic marker includes the lexicalization of a new meaning in everyday speech through its usage as the speech automatism and

the assignment the special function to this marker in a certain communicative context [3].

The group of various discourse markers is formed by the words and phrases which are grammatically parts of speech, and the presence of this term, for the most part, points at the new approach of discourse analysis and constitute the opportunity to investigate relations of discourse more precisely. The words belonging to the group of discourse markers are different parts of speech, however, all of them have the ability to structure the pronounced speech or the written text. The range of pragmatic markers, as it is supposed here, consists of functionally "new" words – pragmatic markers, which have as their sources the full meant already existed lexemes, but for now are related to original words as homonyms. Thus, the class of discourse markers is largely the way of analyzing the text considering the functions of markers which manage it, whereas the group of pragmatic markers, it can be said, actually forms a new independent circle of functional words through their usages as speech automatisms, see examples below:

1. 'vidish/-te' (V, 2, Sing./Plur.) (*you see*) is used to attract the listener's attention to the subject of speech, but not to point at the item that both the speaker and the listener see (e.g., it is used during telephone conversation);

2. 'sejchas-sejchas-sejchas' (*one moment*) or 'minutochu-minutochku' (*wait a minute*) appear in the speech as hesitation pragmatic markers which forces the listener to wait a moment until the word, that is looking for by the speaker, is found.

The distinction between pragmatic and discourse markers is formed by the following points [11], [12]:

a) PMs are used in speech unconsciously, without any reflection, at the level of speech automatisms; DMs are put in text consciously, in order to structure its parts in a certain order;

b) PMs do not have (or have weakened, slightly vanished) lexical and/or grammatical meaning; they are almost completely "agrammatical"; DMs have full lexical meaning and grammatical paradigm;

c) PMs are not content or valuable units of speech, they have only functions; DMs have their own definite meaning as content words;

d) PMs are used essentially only in oral spontaneous speech and cannot be found in written texts (except for oral speech imitations, e.g., in modern plays or movies); DMs are presented both in written and oral texts equally;

e) PMs usually express speakers' attitude to the very process of speech production with all related difficulties being sometimes meta-communicative [13]; DMs always convey only speakers' evaluation of the subject discussed and its characteristics, but not of the text that they produce;

f) PMs are not included in the dictionaries in their functional diversity; DMs are the integral part of traditional lexicography as words, from the one hand, and are the subject of discourse related studies, from the other hand.

The typology of pragmatic markers is discussed in details in the section of presented paper which concerned the annotation of material and the system of tags.

2 Practical Significance of the Annotation of Pragmatic Markers

The results obtained by means of analysis of large corpus material allow clarifying traditional views of communication act using the identifying such discourse units–different types of pragmatic markers–which are uttered in speech in order to solve the particular communicative tasks. With the help of PMs, a speaker explicitly verbalizes his/her communicative intensions, attitude to the addressee, and appeals to the common with his/her interlocutors' perceptual basis. Because of the presence of PMs, the hearer can percept not only truth-conditional, informative level of speech, but also its structural level, as well as can understand how the communication itself functions: the beginning and the end of a speech act or an utterance, the search for words and omissions of lexemes, stressing of the important parts, any disfluencies and call to continue the interaction are marked.

The detailed elaboration of the spontaneous speech pragmatic annotation permits to create the algorithms of automatic checking of the annotation. Approximately each PM has its homonymic analogue which has a full meaning in sentence and is a part of speech, so that the distinction based on hesitation pause after the PM, e.g. 'sejchas', cannot be used since the hesitation break can follow the pronoun 'sejchas', as well as the homonymic PM, too. Each decision about the marking of the PM should be made taking into account the context near PM-"candidate". However, further annotation steps, for sure, will show that some kind of automatism can be presented in the tagging. The ability to implement in the natural language processing system the analysis of functional and structural sides of language, for its part, will contribute to the artificial perceptual basis forming. The modeling of realistic speech dialogues "human–computer/robot/machine" interfaces, that is the most relevant issue in robotics and artificial intelligence development, will be also possible to improve.

The receiving of a full inventory of pragmatic markers of oral speech is also important in such applications as linguodidactics and translation practice. In particular, the introducing of the natural spoken speech materials into textbooks for

the foreign students is essential for training them to understand Russian fluent speech and to avoid plenty of communicative failures. PMs that are used by the native speakers easily and naturally, at the level of speech automatisms, do not prevent to perceive the meaning of a message, and leave beyond the frame of their perceptual field [14]. These markers fall into the perceptual field of foreign speakers and can cause great challenges in communication using a non-native language.

Besides, the typical range of pragmatic markers could be individual for the particular speaker; consequently, this information may be used for the identification of diagnostic features of some age, gender, social or psychological group during conducting linguistic or forensic expertise of oral speech audio recordings.

As one could see, the annotation of the pragmatic markers is required for different linguistic, scientific, and practical needs. This study presents one of the possible ways to organize the process and to develop the methods of the pragmatic annotation that can be applied to analysis of different corpora data.

3 Research Material

The research was carried out on the material from the corpus of Russian everyday speech "One Day of Speech" (ORD), which is one of the most representative resources for the analysis of Russian oral spontaneous dialogic and polylogic speech. The ORD corpus contains 1,250 hours of speech files recorded from 128 informants, which are native speakers of Russian, living in St. Petersburg, and more than 1,000 of their interlocutors, all of them represent various social groups [15, 16]. The records were made using a method of the 24-hours recording of speech day [17] and, after recording, received material were transcribed in the ELAN linguistic annotator. The ELAN files contain several main levels of annotation: transcribed phrase, speaker who pronounced the particular phrase, his/her voice characteristics, events in real life that accompanied the recording, phonetic and phrase commentaries, notes, and episode to which this communicative situation belongs [18].

The pilot subcorpus balanced by gender and age was created for the first annotation of pragmatic markers. The annotation of 12 episodes of corpus speech taken from 12 recordings of different speakers was performed by the group of four annotators independently one from another; total duration amounts 1 hour 46 minutes, 10259 word tokens. For the annotation, additional levels in the ELAN files were made:

- ***PM***, which contains the pragmatic marker in its orthographical form;
- ***Function PM***, that indicates the functions of the PM;
- ***Speaker PM***, which marks the speaker's code;
- ***Comment PM***, that reflect other commentaries connected with the specific PM usage.

4 Development of the System of Tags and Stages of the Annotation

For the annotation, the special system of tags was elaborated that included references to the groups of pragmatic markers already described in the scientific literature [3], [12], [19]. Briefly, for the marker from each group the function manifested in its name is main, but there are plenty of markers that have several functions, i.e., share the common feature of multifunctionality. In the typology of tags below that was developed matching with the system of pragmatic markers itself, the cases of marker polifunctionality are specially commented.

1. APPR – marker-approximator that expresses speaker's uncertainty and hedge:

- *ne znayu // *P vidish' / chego-to Kirill% govorit / chto gips luchshe / yesli (e-e) / tsement bystro vysokhnet / v malen'kikh dyrkakh* ***kak by*** */ yesli tsement bystro vysokhnet / to (:) on ne budet prochnym* [S1];

2. DEICT – deictic marker that points at something vague and consists of 3 elements, two of which are 'vot':

- *nu v obshchem defekt kishki / kogda (e) na nej takoj otrostochek / kak byvaet vot (...) (e-e) v venakh / kak appendiks /* ***vot takoj vot*** *kakoj-to tam* [S130];

3. ZAMEST-PR – replacement marker for the whole set of enumeration or its part:

- *Natasha% / vy uzhe otpustili etogo / () Alekseya%(:) / Maksima% / i* ***vsego prochego*** *? *P vot* [S19];
- *ya govoryu ya togda v devyati tri... tam k devyati pyatnadtsati pridu / poka* ***to syo****...* [S124];

4. ZAMEST-CHR – replacement marker for someone's speech, e.g., 'bla-bla-bla':

- *a / my s toboj zhe byli / pomnish' / Nastya% i Katya%. Aaaa… Kat'ku% ya videla paru raz v universitete / nu / my s nej poskol'ku ne obshchalis' / postoyali / «privet-privet» tam /* ***bla-bla-bla*** [this example is borrowed from the Russian National Corpus];

5. XEN – quotational marker which marks someone else's speech before its appearance in the utterance:

- *nikto poka nichego ne mozhet vnyatnogo skazat' / vse tol'ko razvodyat rukami / (e) i **govoryat** / nu / sochuvstvuyu tipa mol / *P namekayut chto(:) prosto da / oforml... oformlyaj novuyu strakhovku i(:) (...) zhivi spokojno* [S110];

6. MET – meta-communicative marker which fulfills meta-communicative function: the establishment of a contact and understanding between speakers and the speaker's reflection on his/her own speech:

- *nu i Vadik% priezzhaet / *P i oni yemu govoryat **slushaj** chuvak my tebe vsyo otremontirovali / *P tol'ko my tebe koroche (...) (e-e) v bak (...) vmesto(:) (e) dizelya devyanosto vos'moj zalili* [S72];
- *nu Andrej% / togda **vy smotrite** / znachit ya do devyati budu (...) nu (e) telefon vyklyuchu / i otvechat' ne budu / to est' ya prosnus' gde-to v devyat' s kopeechkami / budu uzhe (e) min... vy uzhe v eto vremya budete ekhat'* [S123] (during telephone conversation);

7. NAVIG – navigational marker which serves as structuring device;

- *nu i (...) a do etogo proverili / zheludok vsyo khorosho / a tut polosnaya operatsiya / **vot** eto ya vsyo ... / **vot** eto pervaya chast' Kazani u menya byla normal'naya / a vtoraya chast' (...) **vot** ya vot na etikh samykh zvonkakh nepreryvnykh* [S130] (the marker 'vot' also fulfill the hesitative function here);

8. SEARCH – searching marker that helps the speaker to find the word or expression he/she is looking for:

- *no pri etom b***d' / *P chuvstvuyesh' takoe na***j opustosheniye ! vnutri katarsis chuvstvuyesh' // **kak eto** b***d' () Gracheva% govorila nado // *V ochishcheniye cherez stradaniye* [S15];

9. REFL – reflexive marker which express speaker's reaction to what is said:

- *v itoge my vyzyvali kakogo-to traktorista // *P # khorosho chto nashli vy traktorista // # ugu // *P **ili yeshchyo chego-to takoye** / i koroche vytaskivali Vadika% ottuda // @ ugu* [S72 and W1];

10. RHYTHM – rhythm-forming marker that attaches rhythm to the utterance:

- *vot sejchas uzhe batarei dali / uzhe on bystro vysokhnet // a tak by vot / vot kogda dozhdi shli / **vot** khorosho bylo by zadelat'* [S1];

11. SELFCORR – marker of self-correction:

- *yarkaya solnechnaya pogoda // govorit' mozhno? tak byl yark… ∫ **eto samoe** ∫ byl ∫ iyul'skij den' / vot / nebo bylo chistym / bezoblachnym / solntse ∫ svetilo* (this case is taken from the corpus "Balanced Annotated

Collection of Texts", another corpus of oral speech, created by the group of the same linguists as creators of the ORD-corpus);

12. START – marker of the beginning of an utterance or the process of speech production:

- *ditya moyo / **znachit tak** // *P ta(:)k ? // v etom (...) (m-m) v sentyabre / budet tut vsyo vot tak / *V a v oktyabre / a ... # analogichnaya situatsiya budet na sleduyushchej nedele // # da // @ a ... / a ...* (the marker 'znachit tak' also fulfill the hesitative function here);

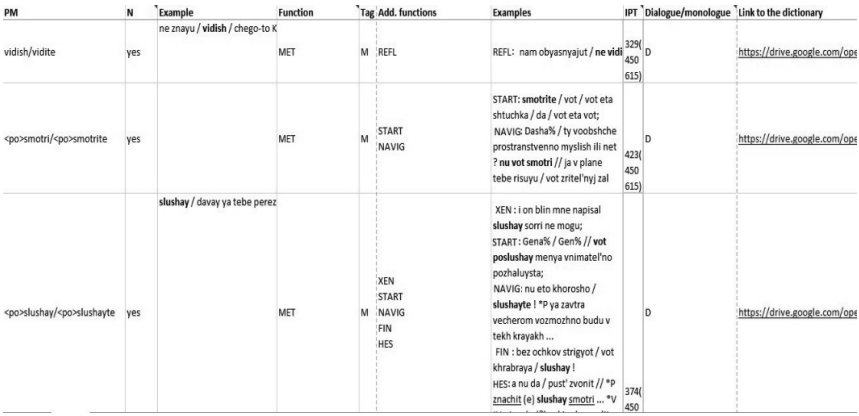
13. FIN – marker of the end of an utterance or the process of speech production:

- *nu ponyatno delo / nu y**ta / a(:) da tebe voobshche / dazhe zakonnyje vykhodnyje mogut ne dat' / da ? **ya dumayu*** [S110] (the marker 'ja duma-ju' also fulfill the hesitative function here);
- *tak / **nu vsyo** / ya ostanavlivayu zapis' / potomu chto eto pustoye / slush-at' eti kliki / vsyo ravno ya nichego bol'she ne skazhu / vse uzhe spyat* [S123];

14. HES – hesitation marker:

- ***nu tam** (...) sil'no deshevle ne bylo / potomu chto ya () zdes' kak by / oni vsyo ravno ekhali* [S103].

The special guideline for the annotators was elaborated. At the first stage of the annotation process, the guideline included the tags consisted of several first letters of particular function (named, as it was showed above), the instructions, such as to write the marker orthographically, to put the tags in the alphabetic order, noting first the main function(-s) of PMs and second the additional function(-s), to separate the repeated markers one from another (do not place them using the hyphen) as well as the description of the process of new level creation in the ELAN program. The possibility to point the new function of a marker was also provided to the annotators. Moreover, before the first try of the annotation, already revealed and described markers were illustrated with an examples from the corpus with an indication of possible functions they can perform. Fig. 1 shows a fragment of the table which was made to help the annotators. The table includes the marker, its structure (one or more words form the marker), examples of usage in speech in the main and additional functions, the tag, items per million value counted in previous researches, the tendency to use it in dialogues or in monologues. In addition, this table contains the link to the document with so-called "described in dictionaries" usages of homonymic to the pragmatic markers expressions. We believed that by producing such table we assisted the annotators to detect the possible pragmatic functions of markers faster and easier.



PM	N	Example	Function	Tag	Add. functions	Examples	IPT	Dialogue/monologue	Link to the dictionary
vidish/vidite	yes	ne znayu / **vidish** / chego-to K	MET	M	REFL	REFL: nam obyasnyajut / **ne vidi**	329(450 615)	D	https://drive.google.com/ope
<po>smotri/<po>smotrite	yes		MET	M	START NAVIG	START: **smotrite** / vot / vot eta shtuchka / da / vot eta vot; NAVIG: Dasha% / ty voobshche prostranstvenno myslish ili net ? **nu vot smotri** // ja v plane tebe risuyu / vot zritel'nyj zal	423(450 615)	D	https://drive.google.com/ope
<po>slushay/<po>slushayte	yes	**slushay** / davay ya tebe perez	MET	M	XEN START NAVIG FIN HES	XEN : i on blin mne napisal **slushay** sorri ne mogu; START: Gena% / Gen% // **vot poslushay** menya vnimatel'no pozhaluysta; NAVIG: nu eto khorosho / **slushayte** ! *P ya zavtra vecherom vozmozhno budu v tekh krayakh ... FIN : bez ochkov strigyot / vot khrabraya / **slushay** ! HES: a nu da / pust' zvonit // *P znachit (e) **slushay** smotri ... *V	374(450	D	https://drive.google.com/ope

Fig. 1. A fragment of the table of described pragmatic markers

After the first stage of the annotation, it turned out that the inter-annotator agreement counted with the help of Kohen's Kappa coefficient (the formula see in [1]) was very low. The best agreement between experts was achieved only for three groups of PMs, i.e., quotational markers, meta-communicative markers, and reflexives. Therefore, the decision to improve the guideline for the annotators was made. Fig. 2 presents a fragment of the table with all possible variants of one marker that can be united by its main type.

This step allows annotating markers automatically and to narrow down the variants to one basic construction. Such variety of grammatical forms reflects the process of pragmaticalization without grammaticalization, as well as the ability of markers to combine with other pragmatic or "meaningless" (functional) components of speech (particles, interjections, conjunctions, etc.), and exists for the all the markers considered in the research: 'eto', 'eto samoje', 'kak jego', 'ne znaju', 'sejchas', 'minutu', 'sekundu', 'tipa', 'vrode', 'kak by', 'takoj', 'bla bla', 'lia lia', 'ili kak eto', 'ili kak jego', 'ili chto yeshchyo' and many others.

For prepare the next stage of the annotation, it was determined, first, not to reduce all the variants of one marker to one basic structure, leaving, during the annotation, the PM in the form in which it was presented in speech, which saved the variety of markers structure; and second, to shorten the list of PMs' functions, so that exclude the most ambiguous cases which revealed total annotators disagreement. Third, the opportunity to list the main and additional functions in a free order was given to the annotators, because of mentioned in the introduction of this paper the multifunctionality of PMs.

A	B	C
Invariant	Variant	Key word for search
prikin'	prikin'	priki...
	prikin'te	
zatseni	zatseni	zatseni
	zatsenite	
glyan'	glyan'	glyan'
	glyan'te	
zamet'	zamet'	zamet'
	zamet'te	
chto yeshchyo	chto yeshchyo	yeshchyo, skazat'
	chto yeshchyo skazat'	
	chego-to yeshchyo khotela skazat'	
	chto yeshchyo b skazat'	
	chto yeshchyo by skazat'	
	chto-to yeshchyo khotel skazat'	
	chego-to yeshchyo khotel skazat'	
	chto-to yeshchyo khotela skazat'	
koroche	koroche	koroche
voobshche	voobshche	voobshche
	voobshche govorya	
sobstvenno	sobstvenno	sobstvenno
tam	tam	tam

Fig. 2. A fragment of the table of variants of pragmatic markers

At the second stage of the annotation process, the new guideline included fewer tags as some of them were grouped (e.g., the group of markers of a boundary (G) unified previous existed start, final and navigational markers), and all the tags were cut to one letter in order to make the annotation process less time-consuming. The annotation of the same files was performed by the same group of annotators independently one from another; they also had been asked to use the new instructions and the system of tags. The analysis of inter-annotator agreement showed the increased level of agreement–up to Kappa=0,51, especially for two annotators who are the authors of presented article [20]. It means that the development of the annotation scheme discussed above, the guideline and the tables of variants improves the results of annotation. The elaborated procedure of the annotation of PMs is supposed to be widely used in the investigations involving the similar methods and data.

However, the process of the annotation cannot be lead without any issues. The human factor and the subjectivity cannot be absolutely removed from the language analysis, but there are certain problems of the annotation that corpus

linguists might deal with. The ways of solution of this kind of annotation problems are described in the next section.

5 Main Annotation Problems of Corpus Material and Ways of Their Solution

During the process of the manually performed annotation of pragmatic markers, the group of annotators, including the authors of this research, confronted several problems involving the functions of PMs, the difference between a PM and a homonymic expressions (see also: [21]), the human factor, the prosodic features of speech, etc. These problems and the possible methods of their solution will be discussed here one by one.

5.1 The Syntagmatic Division of Spontaneous Speech

One of the most important issue was the syntactic and intonation division of speech in syntagmas that cannot be clearly defined in some cases. The addressing of such ambiguity is relevant for the definition of the PM 'vot' functions that performs as a marker of start or final of a phrase or speech part, according to its pre- or postposition:

- *da / poka vot () Marina% ne sde... da / i ne posmotrit i ne otfotografiruyet // *P* ***vot*** *// *P vsyo // pozhalujsta // vsego dobrogo / do svidaniya* [S19];
- *ya sejchas pozvonyu Marine% / i vyyasnyu // delo v tom chto / k vam sobiralas Marina% yekhat' Zhdanova% // ne ne ne ne ne ne // *V Marina% Glukhareva% // *N* ***vot*** */ *P i (:) (e-e)* ***vot*** */ ya vyyasnyu / poyedet ona segodnya ili zavtra k vam* [S19];
- *postoyannye koroche / bunty kakiye-to / sobraniya kakikh-to partij raznykh / politicheskikh / tam vsyakikh // tam b***d' partiya na partiyu / koroche /* ***nu vot*** *// *P zastrelili / odnogo na ulitse / sluchayno // *P (e)* ***vot*** */ *P vtoroj spilsya / a glavnyj geroj / koroche / u nego umerla eta devushka* [S15];
- *moj Seva% byl (...) v techeniye (...) tryokh / chetyryokh dnej v reanimatsii //* ***vo(:)t*** */ sejchas ya yedu / (...) prosto poyedu / net / nu yego uzhe vypisyvayut v chetverg / poyedu povezu / on menya poprosil / chto privezti* [S130].

The pause after the marker means that the topic shift takes its place in the utterance. This unit can be classified as the PM of start due to its position in the beginning of a new phrase. However, it is not defined in these examples, whether the marker attributes to the new topic or discourse fragment itself or the marker closes the previous speech segment with the meaning of conclusion.

The annotation of the start, navigational and final markers caused disagreement at the first stage of the annotation. It is obvious that all these markers share one common function–the marking of a boundary, with the possible change of topic, the communication strategy, the conditions or a manner of speech producing, etc. However, practically, in speech several markers can serve merely in one definite function, e.g., 'znachit tak' for the marking of start or 'vsyo' for the marking of the end of speech. Despite this, the most commonly used markers of this type–'vot' and 'koroche'–tend to appear in different positions in phrases, not having only one preferable place of occurrence. Therefore, the new annotation rules were implemented at the second stage. As a result of the annotation, the receiving of a complete list of markers, as well as their functions, which all the annotators could agree with, the main goal of the researchers was achieved. The variety of "boundary"-tags resulted in inter-annotator disagreement, which showed the disadvantages of tags system. The reduction of tags by clustering them into groups led to making the functions more identifiable. Thus, one tag "G" was produced to unite different tags of boundary markers: "START", "FIN", and "NAVIG". The specifics of each case of boundary PMs will be described during the qualitative analysis of the material after the annotation of all corpus data. Moreover, the distinctive features of different types of boundary PMs are planned to elaborate.

5.2 Pragmaticalization as a Continuing Process

The annotation of pragmatic markers is complicated by the live processes existing in oral spontaneous speech, i.e. grammaticalization and pragmaticalization. Thus, the different degrees of pragmaticalization, a closeness of a unit to the PM class, can be distinguished, e.g.:

- *nu ya sproshu // yesli tsementa ne budet / togda ya gips voz'mu // # v malen'kikh dyrkakh / *P dlya bolshikh dyrok gips ne podkhodit / a () dlya bolshikh dyrok podkhodit tsement // *P* ***ya dumayu*** *// nu ya ne znayu / *P chto takoye bolshaya dyrka // *P v takom-to vot sluchaye* [W1 and S1];
- *nu ponyatno delo / nu y**ta / a(:) da tebe voobshche / dazhe zakonnyje vykhodnyje mogut ne dat / da ?* ***ya dumayu*** *// *P u menya tam podnakopilos' etikh samykh / neispol'zovannogo otpuska / da / poetomu ya i ispol'zuyu* [S110];
- **P kak to tak ona korotkovata nemnozhko poluchilas' // vrode yeshchyo odin shkaf prositsya // *P kholodilnik ne vkhodit a / tak mesto svobodnoye est' // *P* ***ne znayu*** [W1];
- *ponyatno / ya prosto khochu vam skazat' / ya ne ... / vernej sprosit' / snachala dlya nachala / potom uzhe skazat' / *V po povodu etoj programmy (:) /*

*vot ona (...) nastol'ko zamedlyayet rabotu komp'yutera / *P chto vot (e-e) / nu mne prikhodyat gigantskiye fajly / **ya ne znayu** chto tam / eto samoye / no ...* [S19].

It seems that the first two examples shows already pragmaticalized usages of VP 'ja dumaju' that only marks the end of a sentence and do not contribute anything to the content. These PMs also reflect the speaker's hesitations and serve as means of a hedge, as well as the unit 'ne znaju' in the third case. It should be noted that there is a possible interpretation of these markers as not fully pragmaticalized, but only taken a pragmaticalization path ones, that are mostly potential, than real, PMs.

The last phrase is truncated, but by the presence of the hesitation ('eto samoje') we can conclude that the speaker does not know what to say next and how to describe the problems with the computer in more detail. It leads us to the assumption that 'ja ne znaju' in this case is the hesitation PM used in preparing, after all, unsuccessful tries to continue the speech production. However, this construction can be also examined as a meaningful sentence, just left by the speaker and not extended further. Since that, the annotation of such case is ambiguous, from our perspective. The variability of analysis is not only possible, but also necessary for dealing with PMs. Perhaps, the annotation of a wider data allows solving the issue of annotating of such phenomena; the experts have to create the acceptable limits up to which the meaning of a lexeme is identifiable and the unit is still not a marker, otherwise, it should be considered a pragmatic unit having only function in oral discourse.

5.3 Main and Additional Functions of PMs

The dynamic aspect of producing speech causes certain difficulties in function attributions: the problem of determination of the main and the additional functions of PMs and their difference is also complicated by permanent changing the PM place in phrases. For instance, in phrases:

- *nu tam v osnovnom sovetskuyu chital / znayesh literaturu // nashu **tam** / a(:) ! vperyod k kommunizmu !* [S15];
- *nu ya pytayus // no tam zhe kak prosto kak by () konkurentsiya // *P // to est' **kak by** dazhe yesli ya podnimayu ruku / to yeshchyo ne ... // *V nu ya v printsipe pochti na kazhdom podnimayu / no menya prosto ne vsegda sprashivayut* [S27]

is not possible to identify precisely whether the approximation or hesitation is the main function of PMs 'tam' and 'kak by'. The role of this PM in the discourse lies in the fact that they help the speaker to have a little pause in speech structuring and give him/her an opportunity to express the idea approximately,

without further description. To determine which function is predominant seems quite impossible here (see also: [21, 22]).

At the second stage of the annotation, we rejected the difference between the main and the optional functions since the inter-annotator agreement in their annotation was very low. Henceforth, beyond the annotation of all the functional sets of a particular marker, it will be possible to determine the criteria of function domination and increasing prominence.

The tagging of a rhythm-adding function was also uncoordinated and inconsistent. The findings of the investigation [23] shows that there are rhythm-forming markers which organize spontaneous speech into isochronous structures:

- *vot sejchas uzhe batarei dali / uzhe on bystro vysokhnet // a tak by vot / vot kogda dozhdi shli /* ***vot*** *khorosho by bylo zadelat'* [S1];
- *nu i (...) a do etogo proverili / zheludok vsyo khorosho / a tut polosnaya operatsiya / vot eto ya vsyo ... / vot eto pervaya chast' Kazani u menya byla normalnaya / a vtoraya chast' (...) vot ya* ***vot*** *na etikh samykh zvonkakh nepreryvnykh* [S130].

We suppose that in the cases (in bold) the rhythm-forming function is realized. The first PM 'vot' in the first example functions as the boarder-marker, the second operates in the field of hesitation only, the third presumably is a particle for new information actualization, and the last forms the rhythm and the rate of the utterance, which are supported by the repetition of 'vot'. The second case also shows a frequent usage of 'vot', one of which can be regarded as the rhythm-forming PM in the last position. However, it is possible that all these markers are the individual way of hesitating of the particular speaker.

5.4 Chains of Markers or One Marker?

The cases of neighborhood of pragmatic markers are quite frequent in the spontaneous dialogues and monologues. It raises the question of what should be considered as a chain of markers and what–as a new complex PM with another function. D. Verdonik, M. Rojc, and M. Stabej [9] analyze discourse markers in the corpus of Slovenian telephone conversations TURDIS and try to deal with cases of markers collocation, describing the most widespread chain of markers at the beginning of an utterance. We suppose that the PM which forms one intonation unit and fulfills one function is one integral marker, otherwise it is the chain of different markers following one another with a hesitation graduation. However, in case of hesitation PMs it is difficult to decide whether the function is intensifying or actually is equally shared by the sequence of markers:

- *pod triumfalnuyu_arku$ tam koroche // **vot tipa** (...) Kebern% ch... nu(:) rasskazyval // *P ya nachal chitat' / ya tak_skazat'(?) sovsem drugoye prochital / chem chto on mne rasskazyval* [S15] (hesitation and approximation marker(-s));
- *vchera my s na... s Nadey% vykhodim s raboty // *P ona menya prosit / u vas est' tam telefon (e-e) Glukharevoy% ? ya govoryu da // *P nu i **znachit tam** (...) nakhozhu / diktuyu yej* [S19] (boundary, hesitation and approximation marker(-s));
- *tam to delay / **tam kak by tam** zadaniye // chego-to kak-to ustayu bezumno na samom dele // *P prosto voobshche kak by / v printsipe i *P ne to chtoby ya pryamo tut tak umatyvayus // da ? no vot real'no ochen' ustayu* [S27] (hesitation and approximation marker(-s));
- *nikto poka nichego ne mozhet vnyatnogo skazat' / vse tol'ko razvodyat rukami / (e) i govoryat / nu / sochuvstvuyu **tipa mol** / *P namekayut chto(:) prosto da / oforml...* [S110] (approximator or quotational marker and quotational marker 'mol', probably not the PM since it is used in written texts);
- *nu smotrite / *P v poldesyatogo / **tak znachit** smotrite Andrey% / ya tut pogovoril / (...) yeshchyo s lyud'mi / mne rasskazali sleduyushcheye / chto vot eto staraya tak nazyvayemaya* [S123] (hesitation and boundary marker(-s)).

The examples above show one of the most interesting tendency of spontaneous speech, which opposes the principle of language (and speech) economy–the language redundancy. The repeated markers also present a challenge for the annotators given that they may be interpreted as one marker since they have the same function or as two or more repeated markers as words:

- *u vas segodnya prikhod budet // *P tak / **minutochku minutochku** / Gul'% // *P tak / ya sejchas pozvonyu Marine% / i vyyasnyu // delo v tom chto / k vam sobiralas' Marina%* [S19];
- **P **tak tak** / **tak tak tak** / *P kto(?) *P privetik* [S117].

However, the existence of non-one-word markers cannot allow using the constituent criteria–a word equals a PM–during the annotation. To solve the issue "one or more markers" we plan to investigate the frequency of such series of PMs in the speech corpus, which can clarify their language status. At this stage of the annotation, only minimal structures are annotated, thereafter the cases of markers combination will be examined more precisely.

The inversion in Russian is one more problem for the automatic annotation of PMs:

- *(e-e) eto dejstvitel'no tak... poka ne ponyal / tak kak eto mne rasskazyval chelovek / kotoryj nichego ne ponimayet // nu vot **v samom etom** *N /*

*prosto skazal / kak eto est' // poetomu elektriki mestnyje / vot troye / s kem ya pytalsya cherez tret'ye litso svyazatsya / vse otkazalis' / potomu chto oni skazali tak / *V yesli sdelat' vsyo eto vser'yoz / to eto dorogo* [S123].

This issue is solved by the containing the list of the possible PMs variations, even performed automatically by combinatorial algorithms.

6 Conclusion

The annotation of pragmatic markers is still a great challenge for the researchers since this is mainly manual process, difficult to automation, which creates the theoretical and practical issues concerning the understanding and the typology of PMs, the definition of their functions, and the investigation of oral unstructured human discourse. In the article, the process of the first annotation of pragmatic markers of Russian spoken speech was fully described, including two stages of the annotation, advantages and disadvantages of proposed approach to the pragmatic level analysis. The annotation concerned the pilot subcorpus, but the annotated material will be expanded. The presented problems of the annotation allowed us to elaborate the guideline for the annotators and the list of tags in such way that the inter-annotator agreement became higher. We state that the inclusive automatic tagging of PMs in oral speech cannot be performed for now, however, the automatic check of the annotation, after obtaining the full list of PMs' variations, to avoid the human factor of missing markers is necessary. The fuzziness and ambiguity of spontaneous speech are significant issues in the NLP-tasks, and the future research might develop to overcome the multifunctionality of some PMs during the annotation process.

Acknowledgement.

This research was supported by the Russian Science Foundation, project № 18-18-00242 "Pragmatic Markers in Russian Everyday Speech".

References

1. Leech, G.: Adding linguistic annotation. Wynne, M. (ed.). Developing Linguistic Corpora: a Guide to Good Practice. Oxbow Books, Oxford (2005).
2. Archer, D.: Corpus annotation: A welcome addition or an interpretation too far? Tyrkkö, J., Kipiö, M., Nevalainen, T., Rissanen, M. (eds.). Outposts of historical corpus linguistics: from the Helsinki corpus to a proliferation of resources. Studies in variation, contacts and change in English eSeries (2012). URL: http://www.helsinki.fi/varieng/series/volumes/10/archer/.

3. Bogdanova-Beglarian, N. V.: Pragmatemy v ustnoj povsednevnoj rechi: opredelenie ponyatia i obshchaja tipologia [Pragmatems in spoken everyday speech: definition and general typology]. Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia [Perm University Herald. Russian and Foreign Philology], 3(27), 7–20 (2014). (in Russ.).
4. Fraser, B.: What are discourse markers? Journal of Pragmatics, 31(7), 931–952 (1999).
5. Fraser, B.: Commentary pragmatic markers in English. Estudios ingleses de la Universidad Complutense, 5, 115–127 (1997).
6. Schiffrin, D.: Discourse markers. Cambridge University Press, Cambridge (1987).
7. Schourup, L.: Discourse markers. Lingua, 107(3–4), 227–265 (1999).
8. Traugott, E.: The role of the development of discourse markers in a theory of grammaticalization. Paper presented at the 12th International Conference on Historical Linguistics, University of Manchester, August 1995. URL: https://www.researchgate.net/publication/228691469_The_role_of_discourse_markers_in_a_theory_of_grammaticalization.
9. Verdonik, D., Rojc, M., Stabej, M.: Annotating discourse markers in spontaneous speech corpora on an example for the Slovenian language. Language Resources and Evaluation, 41(2), 147–180 (2007).
10. Aijmer, K.: Pragmatic markers in spoken interlanguage. Nordic Journal of English Studies, 3(1), 173–190 (2004).
11. Bogdanova-Beglarian, N. V., Filyasova, Yu. A.: Discourse vs. pragmatic markers: a contrastive terminological study. In: 5th International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2018, SGEM2018 Vienna ART Conference Proceedings, 19-21 March, 2018, vol. 5, pp. 123–130 (2018).
12. Bogdanova-Beglarian, N. V.: O vozmozhnykh kommunikativnykh pomekhakh v mezhkul'turnoj ustnoj kommunikacii [On the possible communicative barriers in intercultural oral communication]. Mir russkogo slova [The World of Russian Word], 3 (2018) (in print). (in Russ.).
13. Zaides, K. D.: Metakommunikativnyje vstavki v russkoj ustnoj spontannoj rechi na rodnom i nerodnom jazyke [Meta-communicative insertions in Russian oral spontaneous speech of native speakers and foreigners]. Kommunikativnyje issledovanija [Communication Studies], 3(9), 19–35 (2016). (in Russ.).
14. Riehakainen, Ye. I.: Vzaimodejstvie kontekstnoj predskazuemosti i chastotnosti v processe vospriyatia spontannoj rechi [The Interaction between Context Predictability and Frequency in the Process of Perception of Spontaneous Speech (on the Material of the Russian Language)], doctorate thesis, St. Petersburg. (2010). (in Russ.).
15. Bogdanova-Beglaryan, N., Asinovskiy, A., Blinova, O., Markasova, Ye., Ryko, A., Sherstinova, T.: Zvukovoj korpus russkogo yazyka: novaja metodologia analiza ustnoj rechi [Sound Corpus of the Russian Language: a new methodology for analyzing the oral speech]. In: Shumska, D., Osga, K. (eds.). Jazyk i metod: Russkij jazyk v lingvisticheskikh issledovaniakh XXI veka [Language and Method: The Russian

Language in the Linguistic Studies of the 21st Century], vol. 2, pp. 357–372, Kraków (2015). (in Russ.).
16. Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Martynenko, G.: Linguistic features and sociolinguistic variability in everyday spoken Russian. In: SPECOM 2017, LNAI, vol. 10458, pp. 503–511. Springer, Cham (2017).
17. Russkij jazyk povsednevnogo obshhenija: osobennosti funkcionirovanija v raznyh social'nyh gruppah [Everyday Russian Language in Different Social Groups]. Collective monograph. Bogdanova-Beglaryan, N. V. (ed.). LAJKA, SPb (2016). (in Russ.).
18. Asinovsky, A., Bogdanova, N., Rusakova, M., Ryko, A., Stepanova, S., Sherstinova, T.: The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: creation principles and annotation. In: Matoušek, V., Mautner, P. (eds.) TSD 2009, LNAI, vol. 57292009, pp. 250–257. Springer, Berlin-Heidelberg (2009).
19. Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Martynenko, G., Baeva, E.: Towards a description of pragmatic markers in Russian everyday speech. In: LNAI, vol. 11096: Speech and Computer. 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18-22, 2018, Proceedings, pp. 42–48. Springer, Leipzig (2018).
20. Bogdanova-Beglarian, N., Blinova, O., Sherstinova, T., Martynenko, G., Zaides, K.: Pragmatic markers in Russian spoken speech: an experience of systematization and annotation for the improvement of NLP tasks. In: Balandin, S., Salmon Cinotti, T., Viola, F., Tyutina, T. (eds.). Proceedings of the 23rd Conference of Open Innovations Association FRUCT. Bologna, Italy, 13–16 November 2018, pp. 69–77. FRUCT Oy, Finland (2018).
21. Crible, L., Cuenca, M.-J.: Discourse markers in speech: characteristics and challenges for corpus annotation. Dialogue and Discourse, 8(2), 149–166 (2017).
22. Crible, L., Zufferey, S.: Using a unified taxonomy to annotate discourse markers in speech and writing. In: Proceedings of the 11th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, London, UK, pp. 14–22 (2015).
23. Bogdanova-Beglarian, N. V., Kisloshchuk, A. I., Sherstinova, T. Ju.: O ritmoobrazujushchej funkcii diskursivnykh jedinic [On rhythm-forming function of discourse markers]. Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologija [Perm University Herald. Russian and Foreign Philology], 2(22), 7–17 (2013). (in Russ.).

The distributive and statistical analysis as a tool to automate the formation of semantic fields (on the example of the linguocultural concept of “empire”)

Victor Zakharov

Saint-Petersburg State University, Universitetskaya emb.,
7-9-11, 199034 Saint-Petersburg, Russia[2]
v.zakharov@spbu.ru

Abstract. The paper presents ongoing results of automatic creation of a semantic field of «empire» in Russian based on distribution and statistical method using corpus data .

A semantic field is a collection of content units covering a certain area of human experience and forming relatively an autonomous microsystem with one or a few centers. The nature of relations within it is mostly named as an association. The idea is to extract from data on syntagmatic collocability a set lexical units connected by semantic paradigmatic relations of various strength using distributional analyses techniques. Nowadays the presence of big corpora and sophisticated algorithms give the possibility and hope to reach a reasonable results.

The first goal of the study is to develop tools and methodology to fill semantic fields by lexical units on the basis of morphologically tagged corpora and special sketch grammar and then to measure the strength of relations between units and to evaluate the method. We were using a corpus system the Sketch Engine that implements the method of distributional statistical analysis. Text material was represented by own topical Russian corpora created from Russian texts of XVIII –XX centuries. In the course of work and to achieve the goal we have solved a number of tasks, have received lists of items filling the semantic space around a concept of “empire” and we are evaluating the method as successive and promising one. At conclusion further steps were identified to clarify the perspective areas of work and to improve the results obtained.

Keywords: Distributive and statistical analysis, Semantic field, Concept of Empire in Russian.

Дистрибутивно-статистический анализ как инструмент автоматизации формирования семантических полей (на примере поля «империя»)

Виктор П. Захаров

[1] Санкт-Петербургский государственный университет,
Университетская наб., 7-9-11, 199034 Санкт-Петербург
v.zakharov@spbu.ru

1 Введение

Понятие «семантическое поле» применяется в лингвистике для обозначения совокупности языковых единиц, объединенных каким-то общим семантическим признаком; имеющих некоторый общий компонент значения. «Поле – совокупность содержательных единиц, покрывающая определенную область человеческого опыта и образующая более или менее автономную микросистему» [5]. В роли таких лексических единиц выступают слова и словосочетания, как нарицательные, так и имена собственные. Сам термин «семантическое поле» имеет различные модификации или синонимы, как-то: лексическое поле, лексико-семантическое поле, функционально-семантическое поле, кластер, тезаурус, онтология и т. п. Каждый из этих терминов по-своему задает тип языковых единиц, входящих в поле и/или тип связи между ними. В основе теории семантических полей лежит представление о существовании в языке некоторых семантических групп, словарный состав которых объединен различными отношениями, как лингвистическими, так и экстралингвистическими, которые представляют собой сложную систему оппозиций.

Семантический признак, лежащий в основе семантического поля, может рассматриваться как некоторая понятийная категория [6, 12]. В трактовке В.Г. Адмони поле характеризуется наличием инвентаря элементов, связанных системными отношениями [1]. В.Г. Адмони усматривает в поле центральную часть – ядро, элементы которого обладают полным набором признаков, определяющих данную группировку, и периферию, элементы которой обладают не всеми, характерными для поля признаками, но мо-

гут иметь и признаки, присущие соседним полям. Поле предполагает непрерывность связей объектов множества, причем на некоторых участках поля создаются области, в которых связи особенно интенсивны, а признаки особенно сильно выражены. Поле предполагает непрерывность связей объектов множества, причем на некоторых участках поля создаются области, в которых связи особенно интенсивны, а признаки особенно сильно выражены. Тогда говорят о лексико-семантических группах – элементарных микрополях, объединяющих слова, обычно относящиеся к одной части речи и наиболее сильно связанные отношением семантической близости. В общем же случае для поля характерна нечеткость границ между частями речи. Теории семантического поля в лингвистике посвящено большое число работ ([2, 4, 8, 9, 20] и др.).

Для полей характерна возможность количественного выражения силы связи между элементами внутри поля, и поэтому эта задача давно является предметом компьютерной лингвистики. Причем в компьютерной лингвистике «семантическое поле» обычно заменяется понятиями «тезаурус» и «онтология».

Задачу моделирования понятийной или терминологической системы можно разбить на две части: выявление системы понятий (лексических идентификаторов понятий) и выявление отношений между ними. В данной работе нас интересует первая задача, а именно, автоматизированное наполнение лексико-семантических полей. Она может решаться «вручную» путем экспликации и формализации профессионального знания, накопленного в системе человеческой деятельности, на основе знаний специалистов и с использованием имеющихся словарей, учебников и других пособий. Этот путь долгий и трудоемкий. Однако поскольку наши знания о мире так или иначе находят отражение в текстах, то можно поставить задачу извлечения системы понятий из текстов. Минимальный набор требований при этом следующий: множество этих автоматически извлеченных понятий должно быть достаточно полным и сами понятия должны быть связаны между собой. Характер связей на этом первом этапе автоматически не устанавливается. В нашем случае можно говорить о принципе когнитивной однородности [14], когда на каждом этапе решается одна задача, в данной работе это выявление множества основных взаимосвязанных понятий вокруг выбранного ядерного элемента (ключевого слова).

В данной работе мы будем говорить о семантической поле “империя”, понимая под ним совокупность пересекающих лексико-семантических групп (слов или словосочетаний), непосредственно или опосредованно связанных по смыслу с концентом “империя”. Выбор данного концепта обусловлен, с одной стороны, его богатым содержанием, с другой стороны,

это содержание по-разному наполняется в разных языках. Эта работа является частью более широкого исследования, посвященного сравнительному анализу наполнения этого поля в русском, английском, чешском и немецком языках.

2 Дистрибутивно-статистический анализ как основа выявления парадигматических отношений

Основная цель данного исследования – выработка методов и адаптация механизмов автоматического выявления набора базовых понятий, относящихся к заданной теме в корпусах русских текстов на основе дистрибутивно-статистических методов. Следующий шаг на этом пути – представление лексических элементов семантического поля “империя” в виде компьютерного тезауруса с количественными характеристиками силы связи между элементами и примерами из корпусов.

Одним из старых и известных методов лингвистического исследования является дистрибутивно-статистический анализ, при котором используется информация о дистрибуции элементов текста и их числовых параметрах. Уже на заре компьютерной лингвистики предпринимались попытки на основе частотной информации о встречаемости лексических единиц в контекстах определенной величины получать по некоторой заданной формуле количественную характеристику их связанности, что впоследствии нашло выражение в методах выявления коллокаций и многословных единиц на основе мер ассоциации.

Одновременно выдвигались идеи распространения этого метода и на парадигматический аспект языка – идеи о том, что парадигматические связи могут выводиться из связей синтагматических [21, 3, 13, 11, 10]. Принцип перехода от изучения текстуальных связей (синтагматических) к системным (парадигматическим) лежит в основе различных дистрибутивно-статистических методик [19, 20. 32]. Считается, что два элемента связаны парадигматически, если оба они текстуально систематически связаны с какими-то третьими элементами. Соответственно, сила парадигматической связи должна возрастать с увеличением *числа* и силы *общих* синтагматических связей [20: 370].

Однако возможности вычислительной техники того времени не позволяли реализовать эти идеи в виде практически работающих алгоритмов и программ. Далее, чтобы можно было говорить о закономерностях любых статистических распределений, нужны очень большие массивы данных [35]. Таковые появились только с созданием больших корпусов текстов.

Одновременно стали появляться и соответствующие программные средства [23, 26, 27, 34, 35].

3 Механизм формирования лексико-семантических групп и полей

Как уже было сказано, парадигматические связи можно вывести из синтагматических. Эта идея была высказана А.Я. Шайкевичем [21] и К.С. Джоунс (K.S. Jones) (PhD thesis) еще в 1960-х гг., но была реализована только сейчас в корпусной лингвистике, где на базе корпусов текстов появилась возможность создать большую базу сочетаемости лексических единиц и на ее основе "вычислять" множество «ближайших соседей» для каждого слова. Математический аппарат для вычисления такого сходства был разработан Д. Лином (D. Lin) [30].

Однако при «переводе» синтагматики в парадигматику также важно также учитывать наличие синтаксической связи между контекстно близкими элементами текста [24]; [31]. Наш подход предполагает описание сочетаемости с помощью лексико-синтаксических шаблонов (иногда их называют лексико-грамматическими или морфологическими шаблонами). В нашем понимании лексико-синтаксический шаблон – это модель языковой конструкции, в которой указываются существенные грамматические характеристики множества лексем, которые входят в языковые выражения, принадлежащие данному классу, и синтаксические условия построения языкового выражения в соответствии с заданным шаблоном (например, учет морфологических признаков лексических единиц в зависимости от контекстных условий).

В системе Sketch Engine [27], которая использовалась нами для формирования корпусов и выявления синтагматических и парадигматических связей, идея лексико-синтаксических шаблонов реализована в форме так называемых «эскизов слов» (word sketch). По определению «эскиз слова» – это одностраничная, автоматически генерируемая на базе корпуса сводка лексическо-грамматической сочетаемости слова, по-другому, сочетаемости в пределах заданных синтаксических формул. Эти «портреты» слов базируются на наборах правил, описывающих грамматические отношения между словами в тексте, которые получили название Word Sketch grammar, или Грамматика шаблонов.

При создании корпуса на основе указанной грамматики и данных морфологической разметки корпуса формируется специальная база данных, представляющая собой триплеты лексико-грамматических отношений.

Статистическая обработка этой базы и вычисляет данные для построения дистрибутивного тезауруса (thesaurus), который для нас является аналогом лексико-семантической группы для заданного термина. Алгоритм вычисления семантических расстояний между элементами группы (кандидатами в группу) и их внутренней кластеризацией описан в [36: sect. 3, 4].

Формализм для грамматики лексико-синтаксических шаблонов использует регулярные выражения над морфологическими тегами. Соответственно, в принципе любой пользователь-лингвист с некоторым опытом и знакомством с вычислительными формализмами может задать свой набор грамматических отношений. Очевидно, что он должен быть при этом знаком с набором тегов и грамматикой языка. Далее эта грамматике лексико-синтаксических шаблонов при создании корпуса подключается к нему, используя стандартный механизм, и тогда функциональные инструменты системы будут формировать результаты, исходя уже именно из этой пользовательской грамматики.

Формализм для грамматики лексико-синтаксических шаблонов основывается прежде всего на линейной последовательности единиц текста и, следовательно, более явно подходит для языков с жестким порядком слов, таких как английский, и менее - для языков со свободным порядком слов, например, для русского, для последнего требуется гораздо более гибкий подход для написания такой грамматики. Дистрибутивно-статистический анализ в нашем исследовании базируется на грамматике лексико-синтаксических шаблонов для русского языка, разработанной М.В. Хохловой [18]. Схожесть дистрибуции слов высчитывается статистически на основе меры ассоциации logDice [33] и с учетом грамматики лексико-синтаксических шаблонов [18, 26, 33, 36].

4 Материал и инструменты исследования

Основной материал исследования – это специально созданный нами совместно с М.В. Хохловой корпус по теме “империя” на основе текстов об империи в русской литературе и культуре конца XVIII – начала XX вв. (105 текстов, 9 млн. токенов). Таким образом подчеркнем, что мы выделяем концепты, существующие в русском языке на протяжении длительного времени и являющиеся отражением русской культуры.

Корпус делится на 4 подкорпуса по хронологическому принципу: 18-ый век (идентификатор подкорпуса XVIII), 1-ая половина 19-го века (XIX-1), 2-ая половина 19-го века (XIX-2) и 20-ый век (XX). Граничные даты подкорпусов выбраны как своего рода «вехи» в осознании понятия империи в

развитии русской общественной мысли. Жанрово-тематическое наполнение – история, литература, публицистика, философия.

Для нашего исследования, как уже говорилось, мы использовали систему Sketch Engine. Главная ее особенность – это наличие специальных средств, реализующих методику дистрибутивно-статистического анализа – «Тезаурус» (построение тезауруса для заданного термина, другими словами, лексико-семантической группы) (см. Рис. 1), «Кластеризация» (группировка единиц тезауруса в кластеры) и «Дифференциация» (выявление сходства и разницы в сочетаемости для пар слов).

Рис. 1. Фрагмент дистрибутивного тезауруса для слова «империя» по подкорпусу 1-ой половины XIX века.

В состав микрополя (лексико-семантической группы) для термина *«империя»* вошли существительные, имеющие с данной лексемой похожую дистрибуцию (входят в одинаковые синтаксические отношения и часто встречаются в одинаковых контекстах): *«держава», «император», «государство», «церковь», «Европа» «христианство», «Рим»,* и др.

Тезаурус в системе Sketch Engine (или, как его называют, дистрибутивный тезаурус) показывает, какие слова имеют схожую дистрибуцию с заданным словом. В этом случае мы говорим о семантической близости или парадигматическом подобии слов. Единицы семантического поля обладают общими синтагматическими и парадигматическими свойствами, что отражает их семантическую близость.

В каждой предметной области значительная часть терминов, как правило, представлена словосочетаниями. Корпусные инструменты предоставляют нам возможность автоматического выявления коллокаций. Другой инструмент системы, выявляющий синтагматические связи между лексическими единицами – это «Коллокации», вычисляющий силу связанности единиц в линейной последовательности на основе 7 мер ассоциации. Но следует добавить, что этот инструмент выявляет не только синтагматические связи, но и парадигматические, выделяющий при достаточно большом «окне» анализа слова одного семантического поля с заданным.

Имеется также инструмент «Лексические портреты», выявляющий коллигации – коллокации в рамках заданных синтаксических моделей (лексико-синтаксических шаблонов). Если инструмент «Коллокации» вычисляет силу связи между словами по всему корпусу, то второй инструмент – в пределах заданной синтаксической формулы (шаблона). В рамках данного исследования грамматика лексико-синтаксических шаблонов использовалась в составе «Тезауруса».

И, наконец, Sketch Engine позволяет выдавать частотные списки лексических единиц, входящих в корпус, которые используются не сами по себе, а как входной материал для контрастивного анализа, когда данные нашего корпуса сравниваются с нейтральным фоновым. Т.е. лексические единицы, относительная частота которых в текстах исследуемого корпуса существенно превосходит частоту этих слов в фоновом неспециализированном корпусе, считаются ключевыми и включаются (могут быть включены) в формируемое семантическое поле.

5 Эксперименты и результаты

5.1 Методика исследования

Была проведена работа в соответствии со следующей методикой.

1) Получение ранжированного списка семантически связанных терминов (минитезаурус) для слова «империя» по каждому из подкорпусов с помощью инструмента «Тезаурус». Максимальное число единиц в гнезде

тезауруса задается равным 40. Каждому термину в каждом минитезаурусе присваивается ранг.

2) Объединение полученных минитезаурусов в один список, представляющий собой лексико-семантическое поле концепта «империя».

3) Выявление пересечения минитезаурусов, в результате чего каждому термину в объединенном списке присваивается «коэффициент стабильности» (k=1, 2, 3, 4, в зависимости от того, в скольких минитезаурусах тот или другой термин встретился). Термины с коэффициентом больше единицы образуют ядро семантического поля. Для этих терминов вычисляются средний и нормированный ранги силы семантической связи с заглавным словом «империя». Нормированный ранг получается умножением среднего ранга на «ранговый коэффициент нормализации»: 1 - для терминов, представленных во всех четырех минитезаурусах, 2 - для терминов из трех минитезаурусов и 3 - для терминов из двух минитезаурусов (Табл. 1). Таким образом, эти коэффициенты понижают ранги терминов, связанных со словом «империя» в большем числе подкорпусов (т. е. в большем числе временных периодов).

4) Ранжирование лексических единиц ядра семантического поля понятия «империя» по нормированному рангу.

5) Ранжирование лексических единиц поля по коэффициенту семантической близости (score).

6) Подсчет относительной частоты (ipm) лексических единиц поля и ранжирование лексических единиц объединенного списка (поля) по относительной частоте.

Table 1. Сводный дистрибутивный тезаурус для слова «империя» (фрагмент)

Подкорпус	Ранг	Lemma	Score	Freq	Коэф-т стабильн.	Средн. ранг	Норм. ранг
XIX-2	1.	австрия	0,216	1014	1		
XIX-2	**36.**	**англия**	**0,131**	**1055**	**2**	**29**	**87**
XVIII	22.	англия	0,095	148	2		
XIX-2	19.	армия	0,149	478	1		
……..	……..	……..	……,,	……	……..	……..	……..
XIX-1	37.	господин	0,085	363	1		
XIX-2	**24.**	**государственность**	**0,143**	**201**	**2**	**19**	**57**
XX	14.	государственность	0,141	143	2		

XX	**1.**	**государство**	**0,245**	**1016**	**4**	**2,25**	**2,25**
XIX-2	2.	государство	0,200	4240	4		
XVIII	3.	государство	0,184	766	4		
XIX-1	3.	государство	0,135	823	4		
XVIII	19.	греция	0,096	135	1		
XX	2.	гуманизм	0,188	195	1		
XVIII	**2.**	**держава**	**0,189**	**424**	**3**	**4,3**	**8,6**
XIX-2	10.	держава	0,165	606	3		
XIX-1	1.	держава	0,143	96	3		
……..	……..	……..	……,,	……	……..	……..	……..
XIX-1	13.	единство	0,108	254	1		
XIX-2	**5.**	**император**	**0,184**	**1381**	**3**	**4**	**8**
XX	5.	император	0,177	295	3		
XIX-1	2.	император	0,141	373	3		
XX	8.	империализм	0,166	297	1		
XX	7.	интеллигенция	0,173	608	1		
……..	….	……..	……..	……..	……..	……..	……..

7) Формирование ранжированного списка коллокатов для слова «империя» для каждого из подкорпусов с помощью инструмента «Коллокации» (Рис. 2).

Максимальное число коллокатов задается равным 50 (выбирается верхняя часть ранжированного списка). Для формирования списка коллокатов используется 4 наиболее эффективных меры: MI.l-og_f, logDice, min. sensitivity и MI, как это было установлено нами в [37]. «Окно» для вычисления коллокаций задается равным от -3 до +3 (три слова влево и три слова вправо от заглавного).

8) Объединение полученных списков коллокатов в один.

9) Выявление пересечения в объединенном списке отдельных списков коллокатов (по подкорпусам) для каждого термина и приписывание им «коэффициент стабильности» (k=1, 2, 3, 4, в зависимости от того, в скольких списках тот или другой термин встретился). Коллокаты (коллокации) с коэффициентом больше единицы добавляются в ядро семантического

Collocation candidates

Page 1 Go Next >

	Cooccurrence count	Candidate count	MI	min. sensitivity	logDice	MI.log_f
P | N **римский**	123	1,166	8.860	0.10549	10.817	42.710
P | N **германский**	80	915	8.589	0.07498	10.369	37.746
P | N **российский**	59	362	9.488	0.05530	10.401	38.847
P | N **австрийский**	45	743	8.060	0.04217	9.670	30.858
P | N **восточный**	42	915	7.660	0.03936	9.439	28.811
P | N **всероссийский**	33	94	10.595	0.03093	9.863	37.362
P | N **падение**	30	379	8.446	0.02812	9.409	29.004
P | N **западный**	38	1,575	6.732	0.02413	8.880	24.663
P | N **византийский**	23	302	8.390	0.02156	9.104	26.665
P | N **предел**	21	761	6.925	0.01968	8.556	21.408
P | N **священный**	20	486	7.502	0.01874	8.721	22.841
P | N **турецкий**	17	503	7.218	0.01593	8.470	20.863
P | N **османский**	15	18	11.842	0.01406	8.823	32.833
P | N **карл**	15	319	7.694	0.01406	8.470	21.334

Рис. 2. Фрагмент списка коллокатов для ключевого слова «империя» (фрагмент)

поля (составные термины-батиграммы). Для этих терминов вычисляются средний и нормированный ранги силы синтагматической связи с заглавным словом. Нормированный ранг получается умножением среднего ранга на «коэффициент нормализации»: 1 – для коллокатов, представленных во всех четырех списках коллокатов, 2 – для коллокатов из трех списков и 3 – для коллокатов из двух списков (Таблица 2).

10) Ранжирование терминов-биграмм ядра семантического поля понятия «империя» по нормированному рангу.

Table 1. Сводный дистрибутивный тезаурус для слова «империя» (фрагмент).

Подкорпус	Ранг	Лемма	Коэф-т стабильн.	Средн ранг	Норм. ранг
XIX-1	**8.**	**австрийский**	**2**	**6,5**	**19,5**
XIX-2	5.	австрийский	2		
XX	21.	англия	1		

……	……	……	……	……	……
XIX-2	**8.**	**византийский**	**4**	**9,25**	**9,25**
XVIII	6.	византийский	4		
XX	12.	византийский	4		
XIX-1	11.	византийский	4		
……	……	……	……	……	……
XIX-1	34.	Габсбурги	1		
XIX-1	**5.**	**германский**	**3**	**9,3**	**18,6**
XIX-2	3.	германский	3		
XX	10.	германский	3		
XVIII	33.	город	1		
……	……	……	……	……	……
XVIII	**8.**	**могущество**	**2**	**13**	**39**
XIX-2	18.	могущество	2		
XIX-2	**37.**	**наполеон**	**2**	**30**	**90**
XX	23.	наполеон	2		
……	……	……	……	……	……
XIX-2	**27.**	**оттоманский**	**2**	**14**	**42**
XVIII	1.	оттоманский	2		
XIX-2	**7.**	**падение**	**3**	**7**	**14**
XVIII	4.	падение	3		
XIX-1	10.	падение	3		
……	……	……	……	……	……

1) Формирование списка ключевых слов для каждого из подкорпусов с помощью инструмента «Word list (Output type: Keywords)». Сопоставимый корпус для этого – ruSkell 1.4 (см. https://www.sketchengine.eu/russian-skell-corpus/).

12) Объединение списков ключевых слов в один и сортировка объединенного списка по «коэффициенту уникальности» (score).

5.2 Результаты исследования

В результате выполнения пп. 1-2 (раздел 5.1) был получен список терминов, представляющий собой наполнение семантического поля «империя» по данным 4 подкорпусов. Этот список включает 112 разных слов (по алфавиту):

Австрия, Англия, армия, варвар, Венгрия, вера, ветер, воинство, война, вопрос, восток, враг, Германия, герой, господин, государственность, государство, Греция, гуманизм, держава, Европа, единство, жар, земля, зло, злодей, император, империализм, интеллигенция, искусство, истина, история, Италия, Казань, католичество, княжение, королевство, культ, культура, Ливония, Литва, литература, луг, мир, мистика, монарх, монархия, мораль, муж, народ, народность, наука, национальность, нация, Новагород, обоз, образование, образованность, общественность, общество, община, орден, освобождение, отдохновение, отец, отечество, перевод, племя, подвиг, покупка, политика, польза, Польша, правительство, право, православие, предание, призвание, присоединение, продажа, произведение, просвещение, процесс, Пруссия, равнина, размышление, революция, религия, республика, Рим, Россия, Русь, Сибирь, социализм, союз, спокойствие, страна, султан, тип, тиран, традиция, Турция, устройство, учреждение, философия, Франция, христианство, царство, церковь, цивилизация, человечество, язык.

В результате выполнения п. 3 было установлено, что из выше приведённого списка 79 слов (79 вхождений из 160) появляется единожды в одном из минитезаурусов, при этом распределение по подкорпусам следующее: XVIII: 32 слова, XIX-1: 16, XIX-2: 14, XX: 17.

33 слова (81 вхождение) появляются в 2, 3 или 4 минитезаурусах, при этом распределение по подкорпусам следующее: XVIII: 8 слов, XIX-1: 24, XIX-1I: 26, XX: 23. Эти 33 слова мы называем ядром семантического поля.

Вот этот список ядра семантического поля «империя» по данным 4 подкорпусов после ранжирования:

а) по алфавиту:

Англия, государственность, государство, держава, Европа, император, искусство, история, культура, литература, мир, монархия, наука, нация, общество, община, политика, правительство, просвещение, революция, религия, Рим, Россия, союз, страна, традиция, учреждение, философия, Франция, христианство, царство, церковь.

б) по нормированному рангу:

государство, император, держава, Европа, царство, церковь, Рим, Франция, христианство, монархия, правительство, страна, общество, филосо-

фия, революция, культура, нация, Россия, литература, государственность, просвещение, религия, мир, искусство, община, политика, история, учреждение, Англия, союз, традиция, наука.

в) по коэффициенту семантической близости (score**):**

держава, государство, общество, союз, государственность, нация, император, политика, культура, страна, община, церковь, царство, христианство, религия, мир, просвещение, правительство, монархия, Европа, философия, Рим, литература, искусство, учреждение, традиция, англия, Франция, история, Россия, революция, наука.

г) по относительной частоте (ipm):

Россия, общество, церковь, мир, история, государство, наука, просвещение, правительство, держава, политика, царство, литература, революция, философия, союз, страна, Европа, община, культура, император, искусство, христианство, нация, учреждение, Англия, религия, Рим, Франция, государственность, традиция, монархия.

Выполнение пп. 7-10 дало следующие результаты.

Всего в сумме было выделено 115 биграмм, в подавляющем большинстве это биграммы типа *Adj+империя, империя+Ngen., N+империи*. Биграммы контактные или разрывные. Еще одна группа слов – термины из парадигматического ряда, уже выявленные инструментом «Тезаурус». Количественные характеристики следующие: 78 биграмм характерны лишь для одного из подкорпусов, 13 – для двух, 10 – для трех и 4 – для четырех.

Ядро синтагматических коллокаций составляют 24 словосочетания:

Российская империя, Византийская империя , империя германской нации, Восточная империя, Священная империя, падение империи, Австрийская империя, Великая империя, пределы империи, Турецкая империя, столица империи, Западная империя, могущество империи, Оттоманская империя, империя Карла, существование империи, восстановление империи, Латинская империя, область империи, империя Рима, империя Наполеона, разрушать империю, эпоха империи.

В результате выполнения пп. 11-12 был сформирован объединенный список ключевых слов, полученный по частотному критерию: значительное превышение относительной частоты в наших подкорпусах по сравнению с нейтральным корпусом. Вот это список.

князь, государь, Булгаков, царь, боярин, Иоанн, посол, отечество, россиянин, воевода, религиозный, Василий, король, войско, императрица, неприятель, Литва, Всеволод, славянин, Димитрий, русский, народ, церковь, бог, митрополит, Ярослав, Киев, крестьянин, хан, духовный, философия, Мстислав, польский, Святослав, Владимир, религия, воин, христианство, народ, государь, Христос, церковный, русский, дух, христианский, пре-

стол, царь, бытие, град, дружина, древний, двор, слава, грамота, откровение, литовский, свобода, император, мысль, учение, вельможа, мысль, святой, вера, народность, свобода, царский, град, пленник, битва, граф, ум, князь, божественный, племя, грек, церковь, вера, Пётр, Франция, просвещение, поляк, душа, человечество, немец, граф, народ, сознание, небо, немецкий, французский, истина, император, Соловьев, Леонтьев, аполлон, Победоносцев, великий, наука, политический, дух, министр, цивилизация, царство, государь, царевич, мир, государство, Европа, смерть, Русь, Польша, православие, София, болгарин, Герцен, Вяземский, общество, воля, воля, римский, идеал, Австрия, мистический, сила, учение, мысль, разум, отечество, Киреевский, дух, истина, цензура, Тютчев, народ, церковь, сочинение, образованность.

Мы можем назвать его периферией нашего семантического поля.

6 Заключение и выводы

Мы видим, что использование корпуса текстов и инструментов системы Sketch Engine позволяет выявлять в автоматизированном режиме синтагматические и парадигматические связи и создавать более адекватное наполнение терминосистемы. Были получены списки слов и словосочетаний, значительно расширяющие имеющиеся лексикографические пособия (Тезаурус РуТез, «Русский семантический словарь», [15: 13], [16: 475], «Константы: словарь русской культуры» [17]). Однако это «статистическое расширение» получилось чрезмерно широким (см. *посол, отечество, воевода, религиозный, религия, воин* и т.д.). Например, встает вопрос, правомерно ли включать в поле «империя» авторов, пишущих о ней (*Герцен, Киреевский, Тютчев* и др.). Очевидно, неправомерно включать в поле «империя» названия народов, населявших империи (*поляк, русский, россиянин, немец*) и соответствующие им прилагательные. Видимо, требуется продолжить эксперименты с другими более жесткими «техническими» параметрами. Очевидно, что эти списки должны быть соотнесены с экспертными знаниями.

Но уже сейчас на основе полученных результатов можно отметить, что по разным параметрам понятие “империя” в разные периоды времени в русской культуре имеет разные коннотации. Так, бросается в глаза существенное отличие текстов 18-го века. Это видно по составу лексики – см. раздел 4.2: из 79 слов тезауруса, «уникальных» только для одного периода, 32 относятся к 18-му веку. Это отличие проявляется и в именах собственных, вошедших в состав поля. И можно вообще сформулировать осторожный вывод, что несмотря на присутствие империи в 18-ом веке в реальности, сам концепт империи в русской культуре в 18-ом веке еще не сложился.

Далее, более глубокий анализ показывает изменение лексического наполнения нашего поля по данным подкорпуса 20-го века. И это при том, что тексты 20-го века в подавляющем большинстве ограничены 1917 годом.

Естественно, после работы автоматизированных механизмов необходимо привлекать экспертов, как для оценки результатов, так и для определения, если требуется, типов связей между элементами поля. Анализ лексики также показывает, что традиционные тезаурусные лексико-семантические отношения для предметных областей в сфере культурно-литературного лексикона манифестируются недостаточно явно. Фактически, большую часть отношений между отобранными базовыми понятиями следует отнести к отношению «ассоциация». Предполагается разработка с привлечением экспертов специально ориентированного набора отношений для данного поля.

Направления дальнейшей работы следующие:

создать единый «ядерный» корпус, сбалансировав разные временные периоды;

создать подкорпус текстов после 1917 года и провести соответствующие эксперименты;

провести эксперименты с другими параметрами инструментов «Тезаурус» и «Коллокации» (в частности, уменьшить количество терминов, включаемых в дистрибутивный тезаурус, и увеличить размер окна выявления коллокаций);

выявить элементы семантического поля (дистрибутивного тезауруса) для терминов, вошедших в ядро поля «империя», т.е. создать тезаурусы (поля) второго уровня, и сформировать объединенный список, по возможности, в виде семантической сети;

провести лингвистическую и культурно-историческую интерпретацию полученных результатов;

разработать или адаптировать программное обеспечение для создания и ведения электронного тезауруса- компьютерного представления поля;

создать электронный тезаурус для семантического поля «империя» с указанием связей между его элементами, с частотными характеристиками и примерами употребления в корпусах.

Благодарности

Исследование поддержано грантом РФФИ № 18-012-00474 «Семантическое поле «империя» в русском, английском и чешском языках» и грантом РФФИ № 17-04-00552-ОГН-А «Параметрическое моделирование лексической системы современного русского литературного языка».

Литература

1. Адмони В. Г. Синтаксис современного немецкого языка: Система отношений и система построения. Л.: Наука, 1973.
2. Апресян Ю.Д. Образ человека по данным языка: Попытка системного описания // ВЯ. 1995, № 1.
3. Арапов М.В. Некоторые принципы построения словаря типа “тезаурус” // НТИ. Сер. 2. 1964. № 4. С. 40–46.
4. Аскольдов С.А. Концепт и слово / Русская словесность. От теории словесности к структуре текста. Антология. М., 1980.
5. Ахманова О.С. Словарь лингвистических терминов. М., 1966.
6. Бондарко А.В. Функциональная грамматика. Л., 1984
7. Васильев Л.М. Современная лингвистическая семантика. М., 1990
8. Вежбицкая А. Понимание культур через посредство ключевых слов М., Языки славянской культуры, 2001.
9. Вежбицкая А. Язык. Культура. Познание. М., Русские словари, 1997.
10. Войскунский В.Г., Захаров В.П., Мордовченко П.Г., Сороколетова Л.И. О некоторых лексико-семантических проблемах в “бестезаурусных” ИПС // Структурная и прикладная лингвистика: Межвузовский сборник. Вып. 2. Л., ЛГУ, 1983. С.170 - 177.
11. Караулов Ю.Н. Лингвистическое конструирование и тезаурус литературного языка. М., 1981.
12. Кобозева И.М. Лингвистическая семантика. М., 2000
13. Пиотровский Р. Г. Текст, машина, человек. Л., 1975.
14. Рубашкин В.Ш. Онтологическая семантика: Знания. Онтологии. Онтологически ориентированные методы информационного анализа текстов. М., 2012.
15. Русский семантический словарь. Т. 2. Москва, Азбуковник. 2002.
16. . Русский семантический словарь. Т. 3. Москва, Азбуковник. 2003.
17. Степанов Ю.С. Константы: Словарь русской культуры. М., 2001.
18. Хохлова М.В. Разработка грамматического модуля русского языка для специализированной системы обработки корпусных данных // Вестник Санкт-Петербургского государственного университета. Серия 9. Филология, востоковедение, журналистика. СПб., 2010. Выпуск 2. С. 162–169.
19. Шайкевич А. Я. Дистрибутивно-статистический анализ текстов. АДД. Л., 1982.
20. Шайкевич А.Я. Дистрибутивно-статистический анализ в семантике // Принципы и методы семантических исследований. М., 1976 . С. 353 378.
21. Шайкевич А.Я. Распределение слов в тексте и выделение семантических полей // Иностранные языки в высшей школе. М.,1963.
22. Щур Г.С. Теория поля в лингвистике. М.-Л., 1974.
23. Blancafort H. Daille B., Gornostay T., Heid U., Méchoulam C., Sharoff S. TTC: Terminology extraction, translation tools and comparable corpora // 14th EURALEX International Congress. 2010. P. 263 268.

24. Gamallo P., Gasperin C., Augustini A., Lopes G. P. Syntactic-Based Methods for Measuring Word Similarity // Text, Speech and Dialogue: Fourth International Conference TSD–2001. LNAI 2166. Springer-Verlag, 2001. P. 116–125.
25. Kilgarriff A., Miloš Husák, Katy McAdam, Michael Rundell and Pavel Rychlý (2008). GDEX: Automatically finding good dictionary examples in a corpus. In Proceedings of the 13th EURALEX International Congress. Spain, July 2008, pp. 425–432.
26. Kilgarriff A., Rychly P. An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments) // Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Czech Republic, June 2007, pp. 41–44.
27. Kilgarriff A., Rychlý P., Jakubíček M., Rundell M. et al., SketchEngine [Computer Software and Informatiom Resource], URL: http://www. sketchengine.co.uk/ Последнее обращение 3.12.2018.
28. Kilgarriff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine // Proceedings of the XIth Euralex International Congress. – Lorient: Universite de Bretagne-Sud, 2004. – P. 105 116.
29. Kilgarriff Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, VítSuchomel. The Sketch Engine: ten years on. In Lexicography 2014? 1(1): 7–36. DOI: 10.1007/s40607-014-0009-9. ISSN 2197-4292.
30. Lin D. Automatic retrieval and clustering of similar words. Proc. COLING-ACL. Montreal: 1998. P. 768-774.
31. Pazienza M., Pennacchiotti M., and Zanzotto F. Terminology extraction: an analysis of linguistic and statistical approaches. Knowledge Mining Series: Studies in Fuzziness and Soft Computing, Springer Verlag, Berlin, 2005. P. 255–279.
32. Pekar V. Linguistic Preprocessing for Distributional Classification of Words // Proceedings of the COLING–04 Workshop on Enhancing and Using Electronic Dictionaries. Geneva: 2004. P. 15–21.
33. Rychlý P. A lexicographer-friendly association score // Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN. Brno. 2008. P. 6–9.
34. Sharoff S. Open-source corpora: Using the net to fish for linguistic data // International journal of corpus linguistics. John Benjamins Publishing Company. 2006. Vol. 11. No. 4. P. 435-462.
35. Smrž P., Rychlý P. Finding Semantically Related Words in Large Corpora // Text, Speech and Dialogue: Fourth International Conference (TSD–2001). LNAI 2166. Springer-Verlag, 2001. P. 108–115.
36. Statistics Used in Sketch Engine. URL: https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine/ (Accessed 03.12.2018).
37. Zakharov V. Comparative Evaluation and Integration of Collocation Extraction Metrics. In: Lecture Notes in Computer Science, vol. 10415 (Text, Speech, and Dialogue - 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings) / K. Ekstein, V. Matousek (Eds.). Springer International Publishing AG 2017. P. 255-262.

Acknowledgement

This work was implemented with financial support of the Russian Foundation for Basic Research, the project No. 18-012-00474 «Semantic field "empire" in Russian, English and Czech» and the project No. 17-04-00552 «Parametric modeling of the lexical system of the modern Russian literary language».

References

1. Admoni, V. G.: Syntax of modern German: The system of the relations and the system of construction, [Sintaksis sovremennogo nemeckogo jazyka: Sistema otnoshenij i sistema postroenija], Leningrad (1973).
2. Apresyan, Yu.D.: The image of a person according to the language: An attempt of the system description, [Obraz cheloveka po dannym jazyka: Popytka sistemnogo opisanija], Linguistics aspects, [Voprosy yazykoznaniya], 1 (1995).
3. Arapov, M.V.: Some principles of creation of the “thesaurus” dictionary NTI Serie 2(4), 40-46 (1964).
4. Askoldov, S.A.: Concept and word, [Koncept i slovo]. Moscow (1980).
5. Akhmanova, O.S.: Dictionary of Linguistic Terminology [Slovar’ lingvisticheskikh terminov]. . Moscow (1966).
6. Bondarko A.V.: Functional grammar, [Funkcional'naja grammatika]. Leningrad (1984).
7. Vasilyev, L.M.: Modern linguistic semantics, [Sovremennaja lingvisticheskaja semantika]/ Moscow (1990).
8. Vezhbitskaya, A.: Understanding of cultures through keywords, [Ponimanie kul’tur cherez posredstvo kljuchevyh slov]. Moscow (2001).
9. Vezhbitskaya, A.: Language. Culture. Knowledge, [Jazyk. Kul’tura. Poznanie]. Moscow (1997).
10. Voyskunsky, V.G., Zakharov, V. P., Mordovchenko, P.G., Sorokoletova, L.I.: About some lexico-semantic problems in «thesauruless» IRS, [O nekotoryh leksiko-semanticheskih problemah v «bestezaurusnyh» IPS]. In: Structural and applied linguistics. [Strukturnaja i prikladnaja lingvistika] vol. 2, pp. 170-177. LGU, Leningrad (1983).
11. Karaulov, Yu.N.: Linguistic designing and thesaurus of the literary language, [Lingvisticheskoe konstruirovanie i tezaurus literaturnogo jazyka]. Moscow (1983).
12. Kobozeva, I.M.: Linguistic semantics, [Lingvisticheskaja semantika]. Moscow (2000).
13. Piotrovsky, R. G.: Text, computer, human, [Tekst, mashina, chelovek]. Leningrad (1975).
14. Rubashkin, V.Sh.:Ontologic semantics [Ontologicheskaja semantika]. Moscow (2012).
15. Russian Sematic Dictionary [Russkiy semanticheskiy slovar’], vol.2, Moscow (2002).

16. Russian Sematic Dictionary [Russkiy semanticheskiy slovar'], vol.3, Moscow (2003).
17. Stepanov, Yu.S.: Constants: Dictionary of the Russian Culture [Konstanty: slovar' russkoy kultury]. Moscow (2002).
18. Khokhlova, M.V.: Development of the grammatical module of Russian for the specialized system of processing of corpus data [Razrabotka grammaticheskogo modulja russkogo jazyka dlja specializirovannoj sistemy obrabotki korpusnyh dannyh], Bulletin of St. Petersburg State University [Vestnik Sankt-Peterburgskogo gosudarstvennogo universiteta], Series 9, Philology, oriental studies, journalism. 2(9), 162-169 (2010).
19. Shaykevich, A. Ya.: Distributive and statistical analysis of texts [Distributivno-statisticheskij analiz tekstov], PhD thesis. Leningrad (1982).
20. Shaykevich, A.Ya.: The distributive and statistical analysis in semantics [Distributivno-statisticheskij analiz v semantike]. In: Principles and methods of semantic researches [Principy i metody semanticheskih issledovanij], pp. 353-378. Moscow (1976).
21. Shaykevich, A.Ya.: Distribution of words in the text and allocation of semantic fields [Raspredelenie slov v tekste i vydelenie semanticheskih polej], In: Foreign languages in higher education. Moscow, 1963.
22. Shchur, G.S.: Field theory in linguistics, [Teorija polja v lingvistike], Moscow-Leningrad (1974).
23. Blancafor,t H. Daille, B., Gornostay, T., Heid, U., Méchoulam, C., Sharoff, S.: TTC: Terminology extraction, translation tools and comparable corpora. In: 14th EURALEX International Congress,. pp. 263-268. EURALEX (2010).
24. Gamallo, P., Gasperin, C., Augustini, A., Lopes, G. P.: Syntactic-Based Methods for Measuring Word Similarity, In: Text, Speech and Dialogue: Fourth International Conference TSD–2001. LNAI 2166, pp. 116–125. Springer-Verlag (2001).
25. Kilgarriff, A., Miloš Husák, Katy McAdam, Michael Rundell and Pavel Rychlý: GDEX: Automatically finding good dictionary examples in a corpus, In: Proceedings of the 13th EURALEX International Congress. Spain, July 2008. pp. 425–432. EURALEX (2008).
26. Kilgarriff, A., Rychly, P.: An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments), In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Czech Republic, June 2007, pp. 41–44. ACL (2007).
27. Kilgarriff, A., Rychlý, P., Jakubíček, M., Rundell, M. et al.: SketchEngine [Computer Software and Informatiom Resource], URL: http://www. sketchengine.co.uk, last accessed 2018/12/03.
28. Kilgarriff, A., Rychly, P., Smrz, P., Tugwel,l D. (2004), The Sketch Engine, In: Proceedings of the XIth Euralex International Congress, pp. 105-116. Lorient: Universite de Bretagne-Sud (2004).
29. Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel: The Sketch Engine: ten years on. In: Lexicography, 1(1), pp. 7–36. DOI: 10.1007/s40607-014-0009-9. ISSN 2197-4292 (2014).

30. Lin, D.: Automatic retrieval and clustering of similar words. In: Proc. COLING-ACL, pp. P. 768-774. Montreal (1998).
31. Pazienza, M., Pennacchiotti, M., and Zanzotto, F.: Terminology extraction: an analysis of linguistic and statistical approaches, In: Knowledge Mining Series: Studies in Fuzziness and Soft Computing, pp. 255–279. Springer Verlag, Berlin (2005),
32. Pekar, V.: Linguistic Preprocessing for Distributional Classification of Words. In: Proceedings of the COLING–04 Workshop on Enhancing and Using Electronic Dictionaries, pp. 15–21. Geneva (2004),
33. Rychlý, P.: A lexicographer-friendly association score, In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN, pp. 6–9. Brno (2008)/
34. Sharoff S.: Open-source corpora: Using the net to fish for linguistic data, In: International journal of corpus linguistics, 4(11). pp. 435-462. John Benjamins Publishing Company (2006).
35. Smrž, P., Rychlý, P.: Finding Semantically Related Words in Large Corpora, In: Text, Speech and Dialogue: Fourth International Conference (TSD–2001), LNAI 2166, pp. 108-115. Springer-Verlag (2001).
36. Statistics Used in Sketch Engine. URL: https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine, last accessed 2018/12/3).
37. Zakharov V.: Comparative Evaluation and Integration of Collocation Extraction Metrics, In: K. Ekstein, V. Matousek (Eds.), Text, Speech, and Dialogue - 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings), LNCS, vol. 10415, pp. 255-262. Springer International Publishing AG, (2017).

Contents

Natalia Bogdanova-Beglarian and Ekaterina Baeva Nonverbal Elements in Everyday Russian Speech: An Attempt at Categorization 5

Vladimir V. Bochkarev Valery D. Solovyev and Anna V. Shevlyakova Analysis of dynamics of the number of syntactic dependencies in Russian and English using Google Books Ngram 18

Rinat Gilmullin, Bulat Khakimov, Ramil Gataullin A Neural Network Approach to Morphological Disambiguation Based on the LSTM Architecture in the National Corpus of the Tatar Language 32

Bagdat Myrzakhmetov and Zhanibek Kozhirbayev Extended Language Modeling experiments for Kazakh 42

Boris Kulik, Alexander Fridman. Roles Contradictions Play in Logical Models of Metaphors and Presuppositions 53

Xenia Naidenova, Sergei Kurbatov and Vjacheslav Ganapol'skii An analysis of plane task text ellipticity and the possibility of ellipses reconstructing based on cognitive modeling geometric objects and actions . . 70

Olga Nevzorova, Damir Mukhamedshin and Alfiya Galieva Named Entity Recognition in Tatar: Corpus-Based Algorithm 86

Elena Sidorova, Natalya Garanina, Irina Kononenko and Alexey Sery Logical-ontological approach to coreference resolution 98

E.V. Sokolova Application of KEA for semantically associated structural units search in a corpus and text summarization 114

Marina Solnyshkina, Valery Solovyev, Vladimir Ivanov, Andrey Danilov Studying Text Complexity in Russian Academic Corpus with Multi-level Annotation 126

A. Vanyushkin and L. Graschenko An overview of the available corpora for evaluation of the automatic keyword extraction algorithms 137

Nadezhda Yarushkina, Aleksey Filippov and Maria Grigoricheva Improving the Quality of Information Retrieval Using Syntactic Analysis of Search Query 151

K.D. Zaides, T.I. Popova, N.V. Bogdanova-Beglarian Pragmatic Markers in the Corpus "One Day of Speech": Approaches to the Annotation 163

Victor Zakharov The distributive and statistical analysis as a tool to automate the formation of semantic fields (on the example of the linguocultural concept of "empire") 182

PROCEEDINGS OF COMPUTATIONAL
MODELS IN LANGUAGE
AND SPEECH WORKSHOP

(CMLS 2018)

Volume 2

В авторской редакции

Подписано в печать: 18.12.2018 г.
Формат 60×84 1/16. Бумага офсетная.
Гарнитура «Таймс». Усл.-печ. л. 11,8 .
Тираж 1 0 экз. Заказ 18В- -2

Отпечатано с готового оригинал-макета ИП Ольшевский В.П.
2 1 -1

ISBN 978-5-9690-0478-8