COGNITIVE STUDIES | ÉTUDES COGNITIVES, 14: 21–31
SOW Publishing House, Warsaw 2014

DOI: 10.11649/cs.2014.003

GANNA KARNAUKH
Ukrainian Lingua-Information Fund, National Academy of Sciences of Ukraine, Kyiv, Ukraine
karnaushka@list.ru

REMOVING GRAMMAR AMBIGUITY OF WORD FORMS BY STATISTICAL METHODS

Abstract

Research is devoted to the study of behavior of linguistic processor at simultaneous application of software supporting functions (taking into account the characteristics of the writing word forms (capital / small letters), punctuation marks in trigrams and location of trigrams within a sentence). The article analyses qualitative quantitative characteristics of the results removing grammatical homonyms of word forms using statistical methods in compliance with requirements. The research is based on the texts of normative legal document.

Keywords: grammatical homonyms, linguistic processor, N-gram, trigram, word form, grammatical meaning, capital / small letters, statistical methods.

Introduction

In the world of linguistics growing awareness leads to the importance of fundamental research of systematic relationships in the triad "information – language – intellect". Evidence of this is the large number of research programs devoted to the study of linguistic intelligence and information systems, connection between language and thinking and cognitive processes, in particular their linguistic aspects. Clearly, new perspectives and approaches concerning the ways of development of linguistic science require solving new challenges: "scientific problem in this area lies in the construction of formal models of linguistic competence, testing and verification of these models and the creation of effective software on their basis..." (Широков, Бугаков & Грязнухіна, 2005). Finally, the need for automated text processing caused the formation of a scientific and technical direction — Natural Language Processing (NLP).

Among the systems of NLP developed in Ukrainian Lingua-Information Fund of National Academy of Sciences of Ukraine (ULIF NASU) widely used are the following: information retrieval systems, systems of grammar annotation, indexing text, machine translation, semantic, statistical analysis and synthesis of speech units, instrumental systems of creating dictionaries of different types (explanatory, translation, spelling, etymological, phraseological, etc.), the management systems of linguistic buildings. Recently, the creation of linguistic buildings is becoming more and more significant, numbering hundreds of millions and even billions of word usage and being the instrument of front language researching both in the phenomenal and contextual aspects. Thus far, there is a problem of identification and qualification of various kinds of linguistic ambiguity, and above all — the removal of grammatical homonyms. The creation of effective automated toolkit for marking buildings (without which the development of Corpus Linguistics in general becomes problematic) depends on successful solution to the last problem.

The removal of grammatical homonyms by statistical methods

We propose a method of eliminating grammatical ambiguity which has three main stages:

- premorphological analysis (separation of a linear sequence of code word in the Ukrainian language (other languages if available), numbers, abbreviations, punctuation marks, and the definition of so-called structural markers — the beginning and end of sentences, paragraphs, texts) is implemented using premorphological analyzer (of program library);

- morphological analysis (determination of word forms grammatical characteristics: the list of elements wf_i with personal settings (v_i, g_i) in length from 0 to n, where wf_i — word form, v_i — part of speech for the word form, g_i — the grammatical meaning of the word form) (See Fig. 1) is provided with the algorithm of lemmatization (Шевченко, Рабулец, & Широков, 2005), based on data of the Grammatical Dictionary of the Ukrainian language[1].

The table of grammatical characteristics of the analyzed word forms[2]

Figure 1

The first triple of word forms in the sentence								
людина			її			життя		
GC	C1	C2	GC	C1	C2	GC	C1	C2
N, f, nom, sg	КИ	(6,1)	P, f, gen, sg	MR	(15,2)	N, n, inan, nom, sg	ЛИ	(13,1)
			P, f, acc, sg	MU	(15,4)	N, n, inan, gen, sg	ЛР	(13,2)
						N, n, inan, acc, sg	ЛВ	(13,4)

[1] Grammatical dictionary of the Ukrainian language [Electronic Resource] by the scientific team of ULIF NASU (Шевченко, Рабулець & Широков, 2005).

[2] Each word form gets the so-called set of grammatical meanings (for all possible homonymous unit grammatical features).

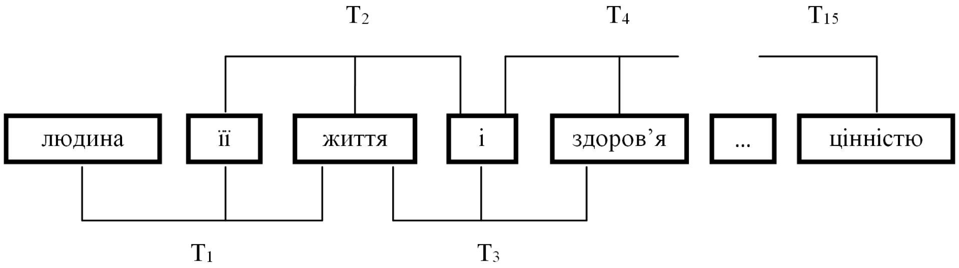
						N, n, inan, voc, sg	ЛК	(13,7)
						N, n, inan, nom, pl	ЛА	(13,8)
						N, n, inan, acc, pl	ЛУ	(13,11)
						N, n, inan, voc, pl	ЛШ	(13,14)

GC — Grammatical characteristics	nom — nominative	sg — singular
C1 — Code 1 (literal)	gen — genetive	pl — plural
C2 — Code 2 (numerical)	acc — accusative	f — feminine
N — noun	voc — vocative	n — neuter
P — pronoun	inan — inanimate	

- statistical analysis (comparison of analyzed word form triples with the text trigrams[3] of training selection[4] — word form lemmatization in the case of absolute trigrams identity or providing a set of possible grammatical characteristics to word forms if trigrams in particular coincide) are related to trigrams base — the array of triples of word forms that contain information on belonging to particular part of speech, trigrams location within a sentence, the grammatical meaning of word forms that it includes, and the presence / absence of punctuation in it. For example:

  Людина, її життя і здоров'я, честь і гідність, недоторканність і безпека визнаються в Україні найвищою соціальною цінністю[5].

Picture 1 Example of sentence division to three word forms, where T is a trigram



[3] In linguistics a combination (group) of three consecutive symbols (Large English-Russian and Russian-Dictionary of English, 2001; New Large English-Russian Dictionary (online version); Universal English-Russian Dictionary, 2011) in this study under the symbol we recognize the word forms (Жигало & Ландэ, 2010; Карнаух, 2011).

[4] The case of manually marked legal texts presented in the form of trigrams.

[5] An example is taken from the Constitution of Ukraine (Chapter 1, p. 3). [Constitution of Ukraine. Law 28.06.1996 Nr. 254k/96-VR / [Parliament of Ukraine]. — Mode of access: `http://zakon2.rada.gov.ua/laws/`.

The table of all possible trigrams analyzed for the three word forms[6]

Figure 2

The first triple of word forms in the sentence			Trigram variants	
			(Type 1[1])	(Type 2[1])
людина	її	життя		
	(15,2)	(13,1)	(6,1) (15,2) (13,1)	(6,1), (15,2) (13,1)
	(15,4)	(13,2)	(6,1) (15,2) (13,2)	(6,1), (15,2) (13,2)
		(13,4)	(6,1) (15,2) (13,4)	(6,1), (15,2) (13,4)
		(13,7)	(6,1) (15,2) (13,7)	(6,1), (15,2) (13,7)
		(13,8)	(6,1) (15,2) (13,8)	(6,1), (15,2) (13,8)
		(13,11)	(6,1) (15,2) (13,11)	(6,1), (15,2) (13,11)
		(13,14)	(6,1) (15,2) (13,14)	(6,1), (15,2) (13,14)
			(6,1) (15,4) (13,1)	(6,1), (15,4) (13,1)
			(6,1) (15,4) (13,2)	(6,1), (15,4) (13,2)
			(6,1) (15,4) (13,4)	(6,1), (15,4) (13,4)
			(6,1) (15,4) (13,7)	(6,1), (15,4) (13,7)
			(6,1) (15,4) (13,8)	(6,1), (15,4) (13,8)
			(6,1) (15,4) (13,11)	(6,1), (15,4) (13,11)

N-gram (or L-gram)[7] is a traditional object of study for Computational Linguistics. The first practical usage was received from the programs determined the texts language written in, and after that from the statistical techniques of automatic translation, eliminating homonyms and others. Besides, N-grams for monolingual linguistic buildings is a commercial product, which has a broad market demand. For example, the N-gram technologies are actively used and promoted by Google and Bing (Жигало & Ландэ, 2010).

Using N-gram method (especially in combination with positional) is quite common in computer linguistics. N-grams are purely descriptive constructs, and the first attempt to introduce them to the field of linguistics was like the usage of the following in an explanatory model role. "It is within association theory of behavior a model of phrase as the Markov process was proposed" (Бузикашвили, Крылова & Самойлов).

Linguistic motivation for going to N-gram is associated with the emergence of descriptive linguistics (Звегинцев, 1959), an arrangement for which, as we see from the title, is the principle of linguistic phenomena fixation combined with theirhigh

[6] The corresponding codes for the analyzed word forms are taken from Fig. 1.

[7] Various studies find both types of writing (which are used as absolute synonyms).

quality theoretical analysis. “Harris’ model of distributive relationship on its lower level is nothing more than N-gram description” (Бузикашвили, Крылова & Самойлов).

However, the meaning of N-grams is quite limited by their applied focus. In any role (at the level of letters, phonemes, word forms, etc.) elimination is an extremely effective tool of one of the problem solution (Бузикашвили, Крылова & Самойлов). Although practical problems in which the rejection is made, is quite different (identifying the part of speech, grammatical forms, the choice of alternatives in the recognition of symbols, possibility of acceptable / probable phoneme, morpheme, word forms in the speech stream), various specific schemes using N-grams (pass forward, going backward or splitting up), but in any case their use is reduced to the imposition of acceptable / probable N-grams to the available data (obtained in the previous step of imposition) and listing of acceptable or most probable chains that are coordinated with such imposition (Бузикашвили, Крылова & Самойлов). Such information is necessary to apply a formal method of removing the ambiguity of language units, which is the basis of proposed program function.

There are many opinions (often quite the opposite) on the usage of the N-gram technology analysis in reference to separate words in a particular language. However, active use of N-grams is observed as a chain of separate words or concepts in a few words (eg, terms) (Саломатина, 2010). Thus, the authors of the article “Concerning the possibility of automating detection of relationships between terms subject area (for example, catalysis)” note: “The term L-gram was probably first introduced by Shannon (Шеннон, 1963) regarding the chain of L arranged following letters of the text in a row, and then it was transferred (not quite correctly) on chains of L arranged following words in a row ($L = 1, 2, ...$)” (Саломатина, 2010). Attractive features of the application of N-grams (represented as chains of separate words) are: their application on texts written in different languages, focus on the removal of arbitrary length units (concepts represented by a word or a few), evaluation of their information content by involving the so-called positional information, the possibility of the formation of patterns to describe N-gram groups and identifying relationships between them (as N-gram spectra contain not only the notion chains, but also indicated (Саломатина, 2010)). In the research we use this variant of N-grams, to be exact — trigrams (the triple of word forms arranged in a row in the text).

This approach to describing the phenomenon of ambiguity takes into consideration the language specificity as an abstract system of linguistic signs and text as a concrete implementation of the system that is the ratio of language / text is observed as contrast of the potency and its realization.

Practical use of the software tool showed that stylistic features, and especially spelling and punctuation of the text, significantly affect the results of eliminating its grammatical ambiguity. The presence of grammar errors in the text (especially spelling) leads to the fact that analyzed unit does not get a set of grammatical characteristics as grammatical vocabulary has no data on it. Moreover, there is a probability of incorrect determination of other word forms within the sentence as a result of a chain effect when identification (valid or not) of one unit affects the labeling neighbors.

Stylistic features of texts also play a significant role. For example, for law texts due to the specific language of law you can develop a number of additional functions, which help to remove unnecessary information from a set of grammatical homonyms of word forms. Due to the fact that grammatical dictionary widely represents Ukrainian language vocabulary (including spoken, outdated, rarely used, proper names along with general) most word forms are defined by program as homonyms.

Considering the features of graphic writing word forms (capital / small letters)

In the written version of the language homonyms, having the similar pronunciation but belonging to different groups on the basis of the ratio of proper / common name are identified at graphic levels. According to the Ukrainian language spelling proper names are written with capital letters (some or all of the words are capitalized in the complex and compound proper names), common names are written with small letters (except when the word is at the beginning of a sentence, a line (in poetry), rubric, direct speech, quotes).

Ukrainian Lingua-Information Fund offered methods of statistical texts analysis, which are used for automatic disambiguation, they take into account several levels of the language: sign, phonetic, lexical, grammatical (Крыгин, 2000), so there is a possibility of distinguishing grapheme for writing (capital / small letter). It gives possibility to formalize the algorithm of removing the ambiguity of word forms, based on the principle of the ratio of proper / common name.

Analysis of the rules for the use of capital / small letters in the Ukrainian language has shown that the word written with small letters always indicates a common name, even if it is a part of complex and compound proper names, denoting a generic concept and having grammatical characteristics similar to word form, which is qualified as a common name (See Fig. 3).

Example: "Київська газета" — the name of the periodical publication and київська газета — newspaper published in Kyiv.

Taking the above into consideration, it can be considered as a correct statement: word form written with a small letter and taking any position in a sentence besides the beginning, can be identified a common name.

This allows formalizing the algorithm for linguistic processor that performs grammatical disambiguation.

The work of algorithm consists of several steps:

1. The choice of word forms.

2. Detection of word forms position within a sentence (position 1 is the first word in a sentence, position 2 — the second, etc.).

3. If linguistic processor captures position 1, ambiguous word form remains unchanged (the ambiguity is not removed), word form occupying the position 2, 3,... is marked in a way that is appropriate to the particular circumstances (in case that grammatical dictionary contains some information about analyzed word form as proper and common name).

Figure 3

Case	Common name		Proper name	
	singular	plural	singular	plural
Nominative	київська газета	київські газети	«Київська газета»	«Київські газети»
Genitive	київської газети	київських газет	«Київської газети»	«Київських газет»
Dative	київській газеті	київським газетам	«Київській газеті»	«Київським газетам»
Accusative	київську газету	київські газети	«Київську газету»	«Київські газети»
Instrumental	київською газетою	київськими газетами	«Київською газетою»	«Київськими газетами»
Locative	київській газеті	київських газетах	«Київській газеті»	«Київських газетах»
Vocative	київська газето	київські газети	«Київська газето»	«Київські газети»

4. Grapheme analysis of word forms: if word form is written with a capital letter, the ambiguity is not removed, if with the small letter — word form is identified as being used to describe the common name (in cases where there is only one variant of grammatical characteristics suitable for common name) the ambiguity of word forms is removed.

Considering the punctuation inside trigrams and its location within a sentence

The practical use of the program shows that the maximum approximation of statistical text's portrait to the analyzed text considerably reduces the number of incorrectly identified word forms. Therefore, there is a need to study the behavior of linguistic processor taking into account the new condition — punctuation marks in trigrams, and to compare the results with previous data. According to this we distinguish two types of trigrams:

Type 1[1]	Type 2[1]
Sequence of three word forms (placed in the text one by one) without regard to punctuation between them	Sequence of three word forms (placed in the text one by one) with regard to punctuation between them
(6,1) (15,2) (13,1)	(6,1), (15,2) (13,1)[8]

[8] Samples of trigrams are taken from Fig. 2.

By the same principle we choose another condition — location of trigrams within a sentence — due to we form these types of trigrams:

Type 1^2: a sequence of three word forms (placed in the text one by one) without regard to trigrams' location in the sentence;

Type 2^2: a sequence of three word forms (placed in the text one by one) with regard to the trigrams' location in the sentence.

Simultaneous use of corrective software functions

The study is conducted on the materials of the text of the Constitution of Ukraine with the amount of 14066 word forms including 2885 unique word forms, 8474 homonymous and 5592 not homonymous word-forms (See Fig. 4).

Analysis of data obtained from the application software, that was aimed to increase the percent accuracy of removal grammatical homonyms with the help of statistical method, shows growth of positive changes and causes the need to investigate the behavior of linguistic processor using these tools simultaneously (See Fig. 5).

General characteristics of the Text 1 and 2[9]

Figure 4

Nr		Text 1	Text 2
1.	Total number of word forms	14066	14066
2.	Number of unique word forms	2886	2886
3.	Total number of marked word forms	9413	9577
4.	Number of marked unique word forms	1879	1747
5.	Number of not homonymous word-forms	4009 (28,50%)	4196 (29,83%)
6.	Number of homonymous word-forms	10057 (71,50%)	9870 (70,17%)
7.	Number of not homonymous unique word-forms	1122 (38,88%)	1181 (40,92%)
8.	Number of homonymous unique word-forms	12944 (61,12%)	12885 (59,08%)

[9] In Text 1 statistical methods for removal were used of grammatical ambiguity of word forms, in text 2 in addition to the general algorithm assistive software functions were used (included writing features of word forms (capital / small letters), punctuation in trigrams, location of trigrams in a sentence).

Results of the removal of the grammatical ambiguity of word forms in the text 1 and 2[10]

Figure 5

Nr		Number of word	
		forms	%
1.	Elimination of incorrect marking (– *)	242	1,72
2.	Elimination of correct marking (+ *)	886	6,30
3.	Incorrect marking (coincidence errors in both texts) (–)	251	1,78
4.	Correct marking of unidentified word forms (*+)	1113	7,91
5.	Incorrect marking of unidentified word forms (*–)	172	1,22
6.	Correct marking of incorrectly identified word forms (–+)	77	0,55
7.	Incorrect marking of correctly identified word forms (+–)	49	0,35
8.	Total number of "positive" changes (amount Nr 1, 4, 6)	1432	10,18
9.	Total number of "negative" changes (amount Nr 2, 5, 7)	1107	7,87
	Total number of "negative" changes (amount Nr 5, 7)	221	1,57

The concept of "positive" / "negative" changes is applied for convenience. The "positive" are those that help to increase the number of correctly identified word forms, as well as decrease incorrectly identified. Therefore, reducing the total number of marked word forms in the text is not a negative effect, because the data includes the errors' elimination (for example, removal of incorrect marking or changing the incorrectly identified into correctly identified word forms).

Using the statistical methods is extremely important to elimination of false marking, as far as the use of trigrams causes the so-called "chain effect" (when analyzed unit affects the others). As a result, one incorrectly identified word form leads to an increase in the number of incorrectly identified items within a sentence. Removal is proper marking of word forms in texts using corrective software functions can be considered as a side effect, which in comparison with the removal of the false marking gets secondary meaning.

The results based on the research show that the behavior of linguistic processor with the simultaneous use of corrective software functions varies considerably. It is important to consider the purpose for removing of the grammatical ambiguity of word forms: general lemmatization of units (as parts of speech) or absolute identification (parts of speech, grammatical meaning), as the removal of the correct marking in most cases is associated with grammatical meaning evaluation of word forms. Trigrams application with regards to punctuation marks and location of word form triples within the sentence is useful (because it facilitates the removal of excess information from a set of grammatical characteristics of word forms) when

[10] Notation conventions: (*+) — comparison of 2 texts (Text 1 and Text 2), positions 1, 2; "*" — unidentified word form, "+" — correctly identified word form, "–" — incorrectly identified word form.

the elimination of grammatical homonyms by statistical methods is considered to be the first stage of electronic data processing (eg, subsequent use of the data with contextual analysis).

Thus, the tendency to reduce the false marking, that is a result of a statistical portrait of the text (in this case punctuation and location of trigrams within a sentence) to the analyzed text, and taking into account the peculiarities of word forms writing (capital / small letters) can increase the positive results of statistical methods for removing grammatical ambiguity of word forms. So software development (algorithms, filters) remains an urgent problem. The software functions must be aimed to increase percent accuracy by elimination of grammatical homonyms and to analyze the behavior of linguistic processor in the application of these tools simultaneously.

References

Apresian, IU. (n.d.). *Novyĭ bol'shoĭ anglo-russkiĭ slovar'*. Retrieved from `http://www.classes.ru/dictionary-english-russian-Apresyan.htm`

Bol'shoĭ anglo-russkiĭ i russko-angliĭskiĭ slovar'. (2001). Akademik.ru. Retrieved from `http://dic.academic.ru/dic.nsf/eng_rus/808936/trigram/`

Buzikashvili, N., Krylova, G., & Samoĭlov, D. (n.d.). *Cognitive Technologies*. Retrieved from `http://cognitive.ru/assets/docs/scienwork/sbornic2/samoilov.doc./`

Karnaukh, G. (2011). Velika i mala litery iak zasib rozriznennia neodnoznachnikh slovoform pri h.avtomatychniĭ dizambiguatsiï. *Problemy gramatyky i leksykologiï ukraïnskoï mowy*, *8*, 26–36.

Kochergan, M. (1980). *Slovo i kontekst*. L'viv: Vyshcha shkola.

Krygin, M. (2000). Tekst na estestvennom iazyke kak ob"ekt statisticheskogo analiza. *Bionika intelektu*, *1*(72), 75–82.

Krygin, M., Shkurko, V., & Afanas'eva, O. (2009). Sniatie grammaticheskoĭ omonimii v tekste s pomoshch'iu statisticheskikh metodov. In V. Shyrokov (Ed.), *Prykladna linhvistyka ta linhvistychni tekhnologiï: MegaLing-2009* (pp. 516–519). Kyïv: Dovira.

Liubchenko, T. (2000). *Leksikohrafichni sistemy gramatychnoho tipu ta ïkh zastosuvannia v zasobakh avtomatychnoho opratsiuvannia movy*. Kyïv: UMIF NANU.

Mazov, N. (1995). N-grammnye metody obrabotki tekstovoĭ informatsii. Ėlektronnaia biblioteka GPNTB Rossii. Retrieved from `http://www.gpntb.ru/win/inter-events/crimea95/report/rep075_r.html`.

Piotrovskiĭ, R. (1979). *Inzhenernaia lingvistika i teoriia iazyka*. Leningrad: Nauka.

Salomatina, N., Gusev, V., Il'ina, L., Kuz'min, A., & Parmon, V. (2010). O vozmozhnosti avtomatizatsii vyiavlenia sviazeĭ mezhdu terminami predmetnoĭ oblasti (na primere kataliza). Konferentsiia «Dialog 2010», doklad 65.
Retrieved from `http://www.dialog-21.ru/digests/dialog2010/materials/html/65.htm`.

Shennon, K. (1963). *Raboty po teorii informatsii i kibernetike*. Moskva: Izd. IL.

Shevchenko, I., Rabulets, A., & Shyrokov, V. (2005). Ėlektronnyĭ grammaticheskiĭ slovar' ukrainskogo iazyka. In V. Shyrokov (Ed.), *Trudy Mezhdunarodnoĭ konferentsii «MegaLing-2005. Prikladnaia lingvistika v poiske novykh puteĭ»* (pp. 124–129). Krym: Meganom.

Shyrokov, V., Bugakov, O., Griaznukhina, T. et al. (2005). *Korpusna linhvistyka*. Kyïv: Dovira.

Universal'nyĭ anglo-russkiĭ slovar'. (2011). Akademik.ru. Retrieved from `http://universal_en_ru.academic.ru/`

ZHigalo, V., & Landė, D. (2010). Statisticheskiĭ onlaĭn-perevodchik InfoStream. In V. Shyrokov (Ed.), *Prykladna linhvistyka ta linhvistychni tekhnologiï: MegaLing2010* (pp. 65-–69). Kyïv: Dovira.

Zvegintsev, V. (1959). Deskriptivnaia lingvistika. In G. Glison (Ed.), *Vvedenie v deskriptivnuiu lingvistiku* (pp. 5–27). Moskva: Izd-vo inostrannoĭ literatury.

Publisher: Institute of Slavic Studies PAS & University of Silesia in Katowice