художника, для религиозного и политического преобразователя всеобщий, истинный идеал не всегда таков, как он признается в его среде» [6, с. 563]. При этом стоит обратить внимание на одну особенность. Как убеждают примеры, советская ораторика в целом не была стандартной, а представляла собой уникальный опыт синтеза общественно-политической, философско-идеологической, производственной, научной, бытовой, антирелигиозной и вместе с тем нравственной проблематики. Все же не подлежит сомнению, что выступления эмигрантов и речи советских риторов этого времени не соответствуют полностью критериям русского риторического идеала.

Список литературы

1. Аристотель. Поэтика. Риторика / Аристотель. – СПб., 2000. – 348 с.

2. Бунин И. А. Миссия русской эмиграции / И. А. Бунин // Русь. – 1924. – 16 февраля – 3 апреля.

3. Ильин И. А. Собрание сочинений : в 10 т. / И. А. Ильин. – М., 1993. – Т. 1. – 398 с.

4. Михальская А. К. Русский Сократ: Лекции по сравнительно-исторической риторике / А. К. Михальская. – М. : Academia, 1996. – 192 с.

5. Рождественский Ю. В. Теория риторики / Ю. В. Рождественский. – М. : Добросвет, 1999. – 482 с.

6. Трубецкой С. Н. О природе человеческого сознания / С. Н. Трубецкой // Сочинения. – М., 1994. – С. 557–563.

FULL DESCRIPTION OF VERB SENSE DISAMBIGUATION IN ENGLISH-AZERBAIJANI MT SYSTEM

Ali Agababa Aliyev

В статье представляется процесс решения многозначности некоторых английских глаголов в англо-азербайджанской системе МП. Описывается процесс создания базы данных формальных признаков решения многозначности глаголов и на основе этих баз разрабатываются алгоритмы решения многозначности в процессе автоматического перевода. Эти базы и алгоритмы представляют начальную аппроксимацию в полном решении многозначности глаголов.

In this paper, full process of verb sense disambiguation is analyzed in the English – Azerbaijani MT system. The creation of database of formal features and algorithms presenting their functioning in translation process is investigated. As initial approximation on the issue of disambiguation of verb senses, the working mechanism is considered as a basic result to the solution of the problem in the English-Azerbaijani MT system. The paper investigates verb sense disambiguation and verb sense generalization issues in Azerbaijani in the language processing while developing English-Azerbaijani MT system. Appropriate methods for disambiguating word senses in Azerbaijani are applied and an initial approach was developed on the investigations.

Ключевые слова: многозначность, машинный перевод, формальный синтаксический анализ, лингвистические алгоритмы, азербайджанский язык.

Key words: ambiguity, machine translation, formal parsing, linguistic algorithms, Azerbaijani.

1. Ambiguity and its effect on the translation quality. The case of linguistic unit having two or more senses, identical in the shade of meaning, is considered as polysemy in natural languages (e.g. the verb *bring* can be translated as *fetch smth,* *give rise, deliver, profit, to raise a question, introduce, convince* and other occurrences and in every case that word expresses an action identical in shade of meaning in depth). But homonymy is the event of a language that encloses the linguistic unit identical in spelling but different in meaning (e.g. *gül-flower (a kind of plant), gül-smile (the imperative form of the verb "gülmək"*) [15; 16]. Both linguistic events – polysemy and homonymy – bear obstacles particularly in the process of formal parsing and the level of the solution to these problems seriously affects the translation quality. The fact that a word form in the source language might have several equivalents (translations) in the target language necessitates the selection of the correct one of those translations. The word sense disambiguation and the part of speech disambiguation are both rather important issues in regard to comprehensibility of the text and the correspondence of the translation of the text to the original alike.

In spite of the fact that, much has been done in scope of the development of applied linguistic technologies for Azerbaijani [1; 4–8], the creation of the algorithmic elimination mechanisms for ambiguous lexical units is still remaining as one of the issues waiting for its solution in the field of machine translation system development for our language. Though one can come across several research work concerning homonymy in Azerbaijani, its types and formal elimination algorithms among these investigations [8; 9], the algorithmic elimination of the polysemy is still remaining as a less investigated field.

Let's have a look at the following examples:

1. Consider the use of the English word *fare* (*yol pulu, sərnişin və s.*) in some contexts:

1) *The conductor gathered the fares midway the destination* (Konduktor yol pullarını mənzil başına qədər olan yolun yarısında yığdı);

2) *The taxi driver got no money from the fare* (Taksi sürücüsü şərnişindən pul almadı).

The word fare has the following translation variants [14]:

fare [fɛə]

1. n

1) yol haqqı, yol pulu;

2) sərnişin;

3) yemək, qida, xörək.

As the result of human identification of a context, the selection of the correct sense bears not so big obstacles (as in the first and second sentences). But in the process of formal translation, if an inappropriate occurrence of a word to the context is selected (e.g. in the first sentence, "yemək" *or* "sərnişin" in place of "yol pulu"), the translation of the sentence will be too wide of the sense of the original.

The disambiguation of an ambiguous word in the process of translation conditions the correspondence of the translation to the original in meaning and consequently becomes one of the factors identifying the quality of the translation. Since the lack of the opportunity of modeling of all the processes going in human brain, the identification of correct sense of all other occurrences of a word is not an overt issue in the process of machine translation. In other words, "to repeat" the capacity of a human to identify the context, that is, the development of algorithms and database of features ensuring the normal functioning of these algorithms to automate this process is still remaining as an actual and open problem for all natural languages without exception [3; 18].

It should be noted that, in formal translation, even statistical calculation does not always help us.

Example:

2. According to the statistics, the English verb *"run"* is mostly used in the meaning of *"to move on foot at a rapid pace"* (in Azerbaijani *"qaçmaq"*). If we translate the sentence *"This young manager is successfully running the bank"* (correct translation *"Bu cavan menecer bankı uğurla idarə edir"*) according to the statistics, then we will get *"bu cavan menecer bankı yaxşı qaçır"*. However, this translation has nothing to do with the original sentence (it expresses no sense in Azerbaijani).

This example proves that the only use of statistics will give rise incorrect results.

Then a question arises whether to show all the possible meanings of a word and leave the selection one of them to the users' will can be considered a way out. The problem is that, presenting all the possible meanings in brackets in the process of translation diminishes the readability of the text to the last extent and turns to an exhausting factor. If there is more than one ambiguous word in a sentence, the presentation of all occurrences of these words will especially make it difficult to comprehend the context.

These examples show to what extent the identification of the correct sense of words is important in the process of formal translation. One of the word groups among ambiguous words is verbs, and this paper is devoted to verb sense disambiguation – to be more precise, the solution to the verb sense disambiguation in the English-Azerbaijani MT system.

2. Existing approaches to Word Sense Disambiguation. Multiple numbers of approaches have been developed to formally disambiguate the ambiguous words: *Naïve Bayesian, Decision List, Nearest Neighbor, Transformation Based Learning, Winnow, Boosting, Naive Bayesian Ensemble etc* [2; 3; 9; 10; 13; 17; 18].

Many of these approaches necessitate the existence of parallel bilingual corpora (a text and its translation into another language), but as there are not such bilingual corpora available for Azerbaijani it is not possible to eliminate this problem using these approaches.

A second group of approaches use the characteristic features – the words used in the near surrounding of the ambiguous word and/or other grammatical features to disambiguate a word [5]. This method of disambiguation is used as a solution to the problem of word sense disambiguation in the English-Azerbaijani MT system.

In this case, the words are analyzed at the level of sentence and formal information sources are used in the sentence. The local surrounding of a polysemous word does much to identify the correct sense of it. On the base of these features, formal rules are created and codified being enclosed in the system.

Example:

3. Consider these two sentences: *I know you* (Mən səni tanıyıram) *və I know this word* (Mən bu sözü bilirəm).

The English verb *to know* is used in both of these sentences. It has to be translated as *"tanımaq"* – *"to be acquainted or familiar with"* in the first case, and *"bilmək"* – *"to have a familiarity or grasp of, as through study or experience"* in the second. For this purpose, it must be entered in the system that, if the poysemous verb is followed by an *animated noun, nouns of place, personal pronouns etc* (this list can be enlarged) then the verb has to be translated in the meaning of *"tanımaq"* – *"to be acquainted or familiar with"* otherwise *"bilmək"* – *"to have a familiarity or grasp of, as through study or experience"*.

3. Formal elimination of ambiguity. The investigations carried out in this field of science show that, some part of the most frequently met errors are directly relevant to polysemy of words in the process of synthesizing sentences in Azerbaijani [8]. The incorrect disambiguation of polysemous words causes the words to differ from the original that is the English sentence in the Azerbaijani translation and as the consequence becomes one of the factors decreasing translation quality in MT systems.

In written sources, the most frequently used English verbs, which have more than two meanings, were primarily selected. For this purpose, English-based sources have been used (http://wordnet.princeton.edu).

Given the fact that, some verbs have multiple translation variants (the second column, in the first table) and every occurrence needs providing formal features, a kind of question arises: is it possible to replace the most identical meanings of one and the same verb with the one that can perform in place of them with special exactness?

The investigations conducted reaffirms to what extent the answer to that question is significant from the point of view of formal analyses. The results of these investigations are shown in the Table I (the number of occurrences of some verbs from English into Azerbaijani).

As seen from the table, after the most identical meanings have been omitted, the capacity of the meaning is reduced approximately 29 %.

Thus, the work on the creation of the database of features for verb sense disambiguation has to be carried out in the following two directions:

1) the reduction of the closest occurrences of polysemous verbs;

2) the selection of features to correctly identify the appropriate meaning.

Consider the following illustrating sentences to notice the possibility of the reduction of the closest meanings.

Let's have a look at the translation variants of the English verb *"subside"*. It has the following translation variants into Azerbaijani.

1) azaltmaq, əksilmək, düşmək;

2) çökmək, enmək, yatmaq (torpaq və s.);

3) sakitləşmək, səngimək (külək, həyəcan və s.) ("Polyglot" electronic dictionary).

It is obviously seen that, the presented meanings have much in general originally. Thus the translations in the second case, as in the first one, express *falling (düşən), weakening (zəifləyən)*, and *descending (enən)* tone of the meaning. (Notice that, in place of "külək səngiyir" *(the wind abates)* "külək zəifləyir" (*the wind weakens)* is also possible formally nevertheless a bit far from the oral speaking style).

Thus, prior to finding formal features for some senses of polysemous verbs, an occurrence that is capable to deliver the meaning of the text to users without misinterpretation has to be selected for some closest meanings. In other words, by limiting the shades of meanings, we broaden the opportunity of making formal the selection process of the shades of meanings.

Sometimes, even though a verb has several translation variants, we replace and enter it the MT system dictionary as one occurrence, since its translations are rather close and give us the chance of substitutability. It prevents not only the unnecessary expansion of the electronic database but considerably decreases the search of formal features for other meanings. After this stage, the number of occurrences (≈ 26 %) of polysemous verbs was remarkably decreased (Table 1. the second column).

The second direction – the work on the identification of the features for formal selection of "context completing meaning" among other occurrences is conducted for each occurrence of every one of the first group verbs (to eliminate this problem, we take verbs as groups to cope with). These features must have characteristics ensuring the unequivocal formal selection of the correct sense of polysemous verbs.

In view of the fact that the sentence is the biggest translation unit In the English-Azerbaijani MT system these features are identified in the frame of sentence. In the process of development of the database for features there was a necessity to analyze a vast amount of sentences which were taken from out of the corpus created in advance. This corpus, in its turn, is made up of several files gathered from different sources. These texts were taken from different internet sites (official chronicle, papers, everyday life, science etc.) to ensure the representativeness of the investigation. Some of the features obtained as the result of this work are shown in the following table.

Table 1

The number of translation variants of some verbs

Verb	Number of translation variants	Reduced number of translation variants
say	6	4
get	19	14
make	10	7
go	21	17
see	8	7
know	6	4
take	26	20
think	9	5
come	7	4
give	15	12
look	3	2
Total	**130**	**96**

Table 2

The database of features for formal elimination of polysemy

Meaning	Code of the feature	Feature explanation	Order
Mean	passive	If the verb is in passive – *nəzərdə tutmaq*	2
Mean	animsubj&vtinf	If the subject is animated and the verb is followed by object + infinitive – *istəmək*	5
Mean		*Istəmək*	3
Mean	clause	If the verb is followed by a clause – *göstər*	5
Mean	inf	Sözdən sonra məsdər – *istəmək*	1
Mean	>[001]	If the verb is followed by another verb – *istəmək*	3
Mean	vtclause	If the verb is followed by an object + a clause – *bildirmək*	7
Mean	animsubj&vtclause	If the subject is an animated noun and the verb is followed by an object + a clause – *nəzərdə tutmaq*	4
Mean	animsubj	If the subject is an animated noun – *nəzərdə tutmaq*	1
Mean		*Bildirmək*	1
Mean		*Nəzərdə tutmaq*	2
Mean	vtinf	If the verb is followed by an object + infinitive – *istəmək*	6
Mean	clause	If the verb is followed by a clause – *bildirmək*	3

It should be noted that, the coded meanings of a verb (We developed a special coding system for this purpose but it is not presented in this paper) is numbered in database. This numbering is not simply implemented from head to foot or vice versa, it is realized by beginning from the meaning with more complex coding to the meaning with simpler one. If the operation is realized beginning from a much simpler code, the symbols of it will overlap with the front part of a much more complex code and the system, accepting it as a proper case, will take the incorrect meaning of the verb as a true one (Sometimes, a simpler code is represented in a much more complex code absolutely the same as in the former. The difference is that, much as the front part is identical, the more complex code includes more symbols. For this reason, much more complex codes have to be analyzed firstly and be passed to simpler ones by stages. However, if the analysis is implemented from a much simpler code to a more complex one, the algorithm will, as mentioned, identify the order appropriate to the code tagged in the sentence and the operation will be completed wrong). In the "from complex to simple" analysis, if a simple coded meaning is needed, it is obvious that, due to the additional symbols of a more complex code it will not be accepted in place of a simpler one. This operation continues till the feature and the coding overlap. From this point of view, the ordering of codes by numbers has rather great effect on the translation quality.

Apart from this it is possible that, none of the features entered will correspond to the context that the word occurs in. Then the meanings with no features – untagged data – will be introduced in accordance with the order in the database, namely, if we can not identify the context of a verb, excluding the first sense, we introduce the meanings of that verb in brackets in "from dominant to less important" order.

4. Schematic illustration of verb sense disambiguation in the English – Azerbaijani MT system. Regarding the above mentioned formal features let's consider the following formal feature schematically.

✓ If the sentence is introduced by words of writing, letter, text etc. type:

As seen from the feature the correct meaning of an ambiguous verb is identified by the formal information the subject encloses. Let's consider a sentence in respect to the case:

Example:

✓ *Today's papers write full page information about the event* (Bugünkü qəzetlərdə hadisə haqqında tam səhifələri ilə yazılır):

In the sentence introduced, it is obviously seen that the arrow from the verb with ambiguity targets the subject. According to the formal rule with the help of the information the subject encloses the correct occurrence of the verb write is identified for the case. Now, let's consider the algorithmic consistency of the process:

Feature 1.

✓ If the sentence is introduced by words of writing, letter, text etc. type:

Algorithm 1

1. Ambiguous verb is defined in the sentence.

2. The subject is taken for formal parsing.

3. The subject is noticed whether it is one of the words of writing, letter, text etc. type.

4. If the subject is introduced by words of writing, letter, text etc. type, the corresponding formal feature entered in the database is taken for translation.

5. Otherwise other formal features are checked.

The main purpose of this algorithm is to define the correct meaning of the verb to the context using the formal information the subject encloses.

Let's consider schematic description of the algorithm introduced above (Fig. 1).

In general the work of the system is carried out in two blocks: the block of analysis and the block of synthesis. In the first case, the morphological analysis, the defination of word phrases syntactical analysis are implemented. The second case encloses the defination of ambiguous word, the connection of suffixes to the words etc. So we can see the whole process of formal translation schematically in the following case:

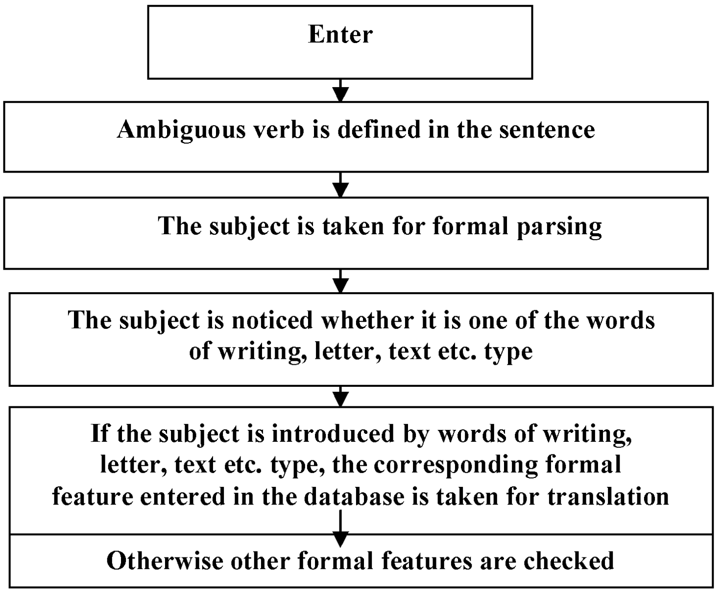
Enter
Ambiguous verb is defined in the sentence
The subject is taken for formal parsing
The subject is noticed whether it is one of the words of writing, letter, text etc. type
If the subject is introduced by words of writing, letter, text etc. type, the corresponding formal feature entered in the database is taken for translation
Otherwise other formal features are checked

Fig. 1. Illustrative outlook of algorithm 1

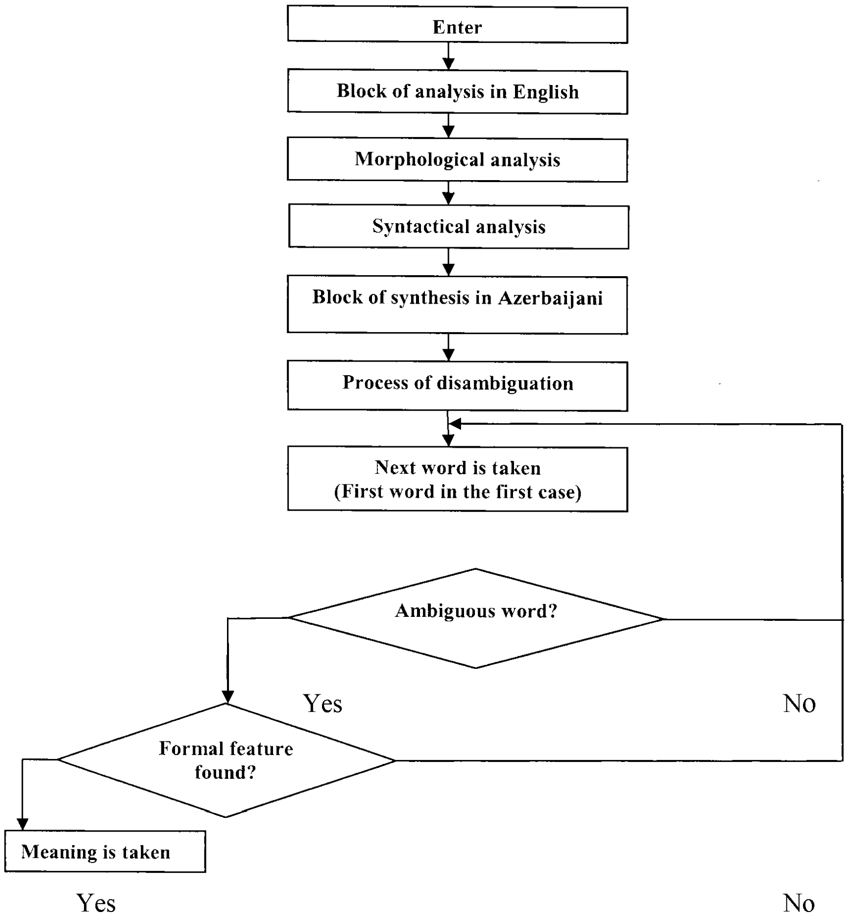
Enter
Block of analysis in English
Morphological analysis
Syntactical analysis
Block of synthesis in Azerbaijani
Process of disambiguation
Next word is taken (First word in the first case)
Ambiguous word?
Yes
No
Formal feature found?
Meaning is taken
Yes
No

Fig. 2. Schematic illustration of formal translation process

Conclusion. The researches in this field for Azerbaijani have been conducted only in the last some years nevertheless the analyses dedicated to the development of Machine Translation Systems cover some decades.

The most frequently used one thousand English verbs were defined for word sense disambiguation; it was defined that the number of meanings of these verbs equals 4852 and occurred that some of these verb meanings absolutely or to a great percentage overlap with another translation variant of the same verb. These meanings were replaced with one meaning best completing overlapping translation variants and finally remained 3899 verb meanings. As the second step, the creation of database of features to correctly select the meaning, their input in the database and coding operations were implemented. Consequently, 387 groups of verb meanings were created and for all these meaning groups special algorithms were developed. These algorithms are considered as the initial approximation to the solution of the problem in the English – Azerbaijani MT system. For better results scientific activities are carried out in the scope of statistics to develop a hybrid method.

References

1. Abbasov A. The use of syntactic and semantic valences of the verb for formal delimitation of verb word phrases / A. Abbasov, A. Fatullayev // Proc. of L&TC'07. – Poznan, Poland, 2007. – P. 468–472.

2. Ahlswede T. E. "Word Sense Disambiguation by Human Informants" / T. E. Ahlswede // Proceedings of the Sixth Midwest Artificial Intelligence and Cognitive Society Conference, Carbondale, Illinois. – April 1995. – P. 73–78.

3. Ahmed F. Arabic-English Word Translation Disambiguation Approach based on Naive Bayesian Classifier / F. Ahmed, A. Nürnberger // Proceedings of the International Multiconference on Computer Science and Information Technology. – Wisla, Poland, 2008. – P. 331–338.

4. Fatullayev A. Formal elimination algorithms of some type homonyms in Azerbaijani sentence / A. Fatullayev, S. Shagavatov // Proceedings of The International Scientific Conference "Problems of Cybernetics and Informatics". – Baku, Azerbaijan, 2006. – P. 108–111.

5. Fatullayev R. Dilmanc is the 1st MT system for Azerbaijani / R. Fatullayev, A. Abbasov, A. Fatullayev // Proc. of SLTC-08. – Stockholm, Sweden. – P. 63–64.

6. Fatullayev R. Peculiarities of the development of the dictionary for the MT System from Azerbaijani / R. Fatullayev, A. Abbasov, A. Fatullayev // Proc. of EAMT-08. – Hamburg, Germany. – P. 35–40.

7. Fatullayev R. Set of active suffix chains and its role in development of MT system for Azerbaycani / R. Fatullayev, A. Abbasov, A. Fatullayev // Proc. of IMCSIT-08. – Wisla, Poland. – P. 363–368.

8. Fatullayev R. Statistical analysis of the factors influencing the translation quality of the Dilmanc MT system / R. Fatullayev, S. Mammadova, A. Fatullayev // Proc. of the International Conference on Problems of Cybernetics and Informatics (PCI-2008). – Baku, 2008. – P. 96–99.

9. Gale K. C. A Method for Disambiguating Word Senses in a Large Corpus / K. C. Gale, D. Yarowsky // Computers and Humanities. – 1992. – Vol. 26. – P. 415–439.

10. Gerard E. Boosting applied to word sense disambiguation / E. Gerard, L. Marquez, G. Rigau // Proceedings of the 12th European Conference on Machine Learning (ECML). – Barcelona, Spain, 2000. – P. 129–141.

11. Hasanov H. A. Dictionary of homonyms of Modern Azerbaijani / H. A. Hasanov. – Baku : Maarif, 1981. – P. 121.

12. Mahmudov M. Metnlerin formal tehlili sistemi (Formal processing system of texts) / M. Mahmudov. – Baku : Elm, 2002.

13. Pedersen T. A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation / T. Pedersen // Proceedings of the First Annual Meeting

of the North American Chapter of the Association for Computational Linguistics, Seattle, WA, 2000. – P. 63–69.

14. Turksever (Musayev) O. İ. English-Azerbaijani dictionary / O. İ. Turksever (Musayev) [et al.]. – Baku : Qismet, 2003. – P. 1673.

15. Wikipedia, the free encyclopedia. – Access mode: http://en.wikipedia.org/wiki /Ambiguity, free. – Title from screen. – English.

16. Wikipedia, the free encyclopedia. – Access mode: http://en.wikipedia.org/wiki /Word_sense_disambiguation, free. – Title from screen. – English.

17. William G. A. A method for disambiguating word senses in a large corpus / G. A. William, C. W. Kenneth, D. Yarowsky // Computers and the Humanities. – 1993. – Vol. 26. P. 415–439.

18. Yarowsky D. Decision Lists for Lexical Ambiguity Resolution:Application to Accent Restoration in Spanish and French / D. Yarowsky // Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, 1994. – P. 88–95.