Studies and Materials in Applied Computer Science, Vol. 3, No. 5, 2011
pp.25-30

USING AN INFLECTION DICTIONARY FOR A DESIGN OF A VISUAL DICTIONARY OF VERBS AND THEIR ATTRIBUTES

Krzysztof Kluza

AGH University of Science and Technology in Kraków
al. Mickiewicza 30, 30-059 Kraków
e-mail: kluza@agh.edu.pl

Izabela Gatkowska

Jagiellonian University in Kraków
ul. Gołębia 24, 31-007 Kraków
e-mail: izabela.gatkowska@uj.edu.pl

Abstarct: *Authors of this paper present a development of a visual web dictionary based on automatic extraction of verbs and their attributes from texts using the inflection dictionary of Polish language. Entries of the designed dictionary are based on triples: subject (agent) + predicate (action) + object/tool. The paper describes a prototype implementation of such a dictionary. In the paper the background for the research is presented as well as the considered problem is outlined. Moreover, the structure of the inflection dictionary of Polish language is described along with the classification of computer dictionaries. In the implementation part of the paper, the architecture of the developed web dictionary and the extraction algorithm are described.*

Keywords: *web dictionaries, visual dictionary, inflection dictionary of Polish language*

1. INTRODUCTION

Dictionaries of different kind have been developed for many centuries. A new branch of linguistics developed in that period - lexicography – describing dictionary development practice and theory [8].
Currently we deal with extraordinary progress in that field. This progress manifests both in the increased number of published dictionaries, constantly increasing automation of the material collection and processing as well as improved dictionary entries description [8].
Internet has become one of the most popular media nowadays. More and more authors, also of scientific studies, use electronic references (sources). Completely new types of dictionaries have been developed in recent years: electronic dictionaries, which are often multimedia and web dictionaries. As a result of diversity of available web dictionaries as well as quick and easy access to them, more and more people use then as potential source of information. However, expansion of modern solutions can be seen mainly in foreign dictionaries available in English. Whereas such diversity is not seen among Polish dictionaries.
Authors of this paper present preliminary results of the research on the development of the visual verb dictionary using Inflection dictionary of Polish language [4].
In order to present the area of research, main classification of electronic dictionaries was provided in the paper herein. Moreover, operation of the Inflection dictionary of Polish language used to develop the dictionary of verbs and their attributes was briefly presented. Main part of the paper includes description of the algorithm of Polish web dictionary of verbs and their attributes as well as presentation of the prototype implementation of the dictionary.

2. MOTIVATION

One says that one picture is worth a thousand words. This is to give better picture of the quantity of information that can be presented in graphic form. Although visual dictionaries are more and more popular, particularly in the English language part of the Internet, one can observe absence of Polish equivalents of such dictionaries.

The project of Polish visual dictionary of verbs presented in this paper, can be, in its final version, useful both for those who learn Polish as a foreign language (as self-study resource) and for those who study Polish as a mother tongue (as a verbal expression aid in writing essays). Moreover, for IT specialists dealing with natural language processing, such dictionary would be particularly useful if the dictionary was provided with an appropriate programming interface.

3. SUBJECT OF THE RESEARCH

Research described in the paper herein concerns the project of the visual web dictionary of verbs and their attributes. More precisely, the issue concerns fully automatic computerized extraction of information from a text, in form of the following triples:

- subject (agent)
- predicate/action,
- object/tool.

The main tool used in the extraction process is the Inflection dictionary of Polish language, described in more detail further in this section.
Polish language is commonly considered as a difficult and complicated. This results from the fact that most of words are subject to inflection, which makes Polish an inflecting language. To properly extract information from a text, we need an inflection dictionary.
The Inflection dictionary of Polish language, developed between 1996 and 2001, includes nearly 120 thousand entries and over 440 inflection patterns. Each entry is an exhaustive and complete collection of inflection patterns of a given word, including its inflexional description. The dictionary was developed as a result of joint research performed in the Department of Information Technology of AGH and the Department of Computational Linguistics of the Jagiellonian University under supervision of Prof. Wiesław Lubaszewski [4]. Word inflexion patters are stored in the dictionary in form of inflexion pattern tree, the fragment of which is presented in the figure 1.

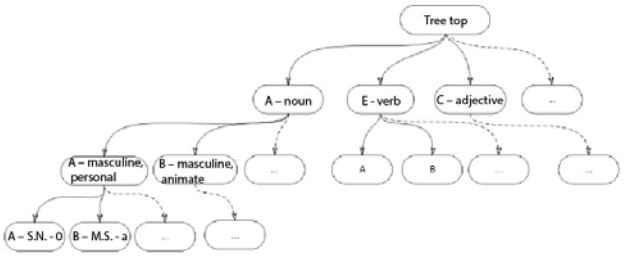

Figure. 1 Fragment of the top part of the word classification system.

Letters of the consecutive nodes of the tree make up the label identifying a given inflexion pattern [3]. For example, from all outermost left branches of the tree we obtain the label AAAAAA, which indicates that the subject-matter word is:

- a noun,
- gender: masculine, personal,
- the singular nominative has zero suffix: S.N. – 0,
- the plural nominative has -owie suffix: P.N. – owie,
- the singular vocative has –e suffix: S.V. - -e,
- the singular dative has –owi suffix: S.D. – owi.

Presented label corresponds to such words like e.g. poseł (member of parliament), bohater (hero), kawaler (bachelor), inżynier (engineer) and the like. The description of the pattern, i.e. also the label, is complete up to the leaf (node), from which it is impossible to pass to further nodes. Each leaf of the tree is provided with an appropriate set of suffixes, comprising of suffixes assumed by words subject to inflection. For the aforementioned label AAAAAA, the suffix vector assumes the following form:

- Nominative, singular -0,
- Genitive, singular - a,
- Dative, singular -owi,
- Accusative, singular - a,
- Instrumental, singular -em,
- Locative, singular -e,
- Vocative, singular -e,
- Nominative, plural -owie,
- Genitive, plural -ów,
- Dative, plural -om,
- Accusative, plural -ów,
- Instrumental, plural -ami,
- Locative, plural-ach,

- Vocative, plural -owie.

In the discussed research, the algorithm of relation extraction from text was created and the dictionary prototype was implemented using the Inflection dictionary of Polish language.

4. TYPOLOGY OF COMPUTER DICTIONARIES

During development of the dictionary, it is worth to take note of its position in the dictionary classification system. Dictionaries can be classified according to different criteria. A detailed typology of dictionaries can be found in monograph [8].
One of the important criteria, modern dictionaries can be classified by, is their form. According to that criterion, dictionaries can be divided into conventional (paper) and electronic (computer) ones. The later group is relatively new type of dictionaries. Although first electronic dictionaries of Polish language [1] designed as an aid not only for native speakers but also for foreigners occurred as early as at the end of the 20th century, it was not until last decade when revival in that field has been observed (the review of selected web dictionaries of Polish language is provided in the publication [2]). However, quantity and diversity of those dictionaries is not equal to English language publications even today.
It is worth to specify it more precisely, that the term electronic dictionary is much wider and it can apply to different types of dictionaries, such as [8]:

- dictionary modules, e.g. in text editors or OCR software,
- computer software on CD/DVD disks,
- dictionary applications available on the Internet,
- electronic equipment designed for translation from/to a foreign language,
- predictive text systems, e.g. T9 for SMS messages.

To differentiate the above types of electronic dictionaries from each other, dictionaries from point 2 are identified as computer dictionaries, while those from point 3 as web dictionaries. Moreover, as regards their functioning, dictionaries listed in point 3 are called online dictionaries, whereas the remaining as offline dictionaries. From the point of view of that classification, the dictionary being developed is a Web dictionary available online.

5. ALGORITHMS FOR EXTRACTING RELATIONS FROM TEXTS

Automatic analysis of compound sentences is a very complex task that requires use of machines capable of handling problems of high computational complexity. Studies on automatic syntax analysis of sentences have been performed in different Polish scientific centres. Information on their results are provided, e.g. in publications [5, 6, 7].
As even rough syntax analysis considerably exceeds the scope of research of the authors of this paper, they focused on development of the algorithm executable using The Inflection dictionary of Polish language only.
To allow quick implementation of the algorithm for determining verb arguments based on the verb use in the sentence, the form of considered sentences was limited. The general chart of the sentence processing algorithm is presented in figure 2.
To enable detection of the agent, verb and object (tool/instrument), the type of processed sentences was limited to those in third person, including explicit subject. Although it seems to be significant limitation, most of texts are in third person.

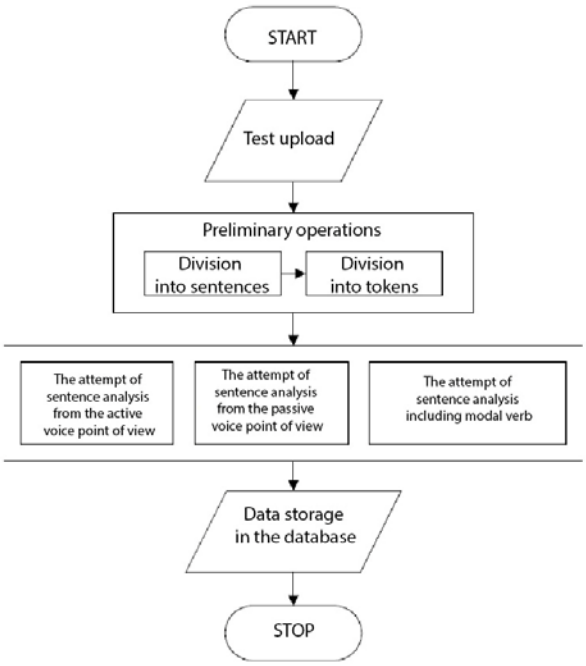

Figure. 2 General chart of the text processing algorithm.

The main reason for excluding first person sentences from the analysis is the problem with determining the subject of

the sentence. Even if there is an explicit subject, like in the sentence Ja wypiłem mleko (I drank milk), we are unable to determine who or what is "I" only on the basis of that sentence (this results from the context only and it may just as well be a man, a cat or even a personalised object).
The analysed sentence can be written in any tense. The analysis takes into account compliance of the number and person between the subject and the predicate, e.g. in case of the following sentences:

- Piekarz piecze chleb (A baker bakes a bread).
- Piekarz piekł chleb (A baker baked a bread).
- Piekarz będzie piekł chleb (A baker will bake a bread).

the first part of the analysis concerns compliance of the person, number and gender for the noun piekarz (baker) and the verb piec (bake). The second part of the analysis concerns the object. Depending on the case, the object is classified as:

- the object (accusative), e.g. pożycza książkę (lends a book),
- instrument/tool (instrumental), e.g. gra na trąbce (plays a trumpet),
- (dative), e.g. daje książkę koleżance (gives a book to a fiend).

If a sentence assumes the following form:

- Piekarz będzie piekł chleb (A baker will bake a bread).

then that sentence will be analysed in the modal verb phase of the analysis, described further in this paper.
The analysed sentence can be written both in active and in passive voice. Due to significant differences in the structure of active and passive voice sentences, passive voice sentences must be analysed separately. For instance, for sentences

- Piekarz piecze chleb (A baker bakes a bread).
- Chleb jest pieczony przez piekarza (A bread is baked by a baker).

From the human point of view, the verb piec (bake) has two arguments in both sentences: the agent piekarz (baker) and the object chleb (a bread). However, the indirect object of the active voice sentence becomes the subject of the passive voice sentence, while the agent is expressed using the structure przez (by) + dative (as in the above example) or using the instrumental case (e.g. when the agent is a tool or an impersonal force: Jan został uderzony kamieniem. (John was hit with a stone) or Jan został porażony piorunem.(John was hit by a thunderbolt)). Hence, during the passive voice sentence analysis, its components are determined as follows:

- the potential object is the subject of the sentence: in nominative and is compliant with the verb,
- potential agent is expressed in dative following the word przez (by) or in instrumental case,
- while the potential verb is expressed by participle following the auxiliary verb być (be) or zostać (become)

Although the passive voice sentence is analysed using the constructed algorithm, then from the computer analysis point of view, the passive voice may turn out to be little useful as the argument expressing the agent is often omitted there, like in sentence Ten piękny obraz został namalowany w 2010 roku (That beautiful painting was painted in 2010). In that sentence, the agent is not necessary at all. As it is known from the common sense, the agent is some painter but it is impossible to extract that information directly from this text.

The developed algorithm is able to properly extract information from sentences of any mood. In case of indicative and imperative mood, the detection algorithm is the same as in the following sample sentences:

- Ten mężczyzna kupi mleko (This Man shall buy milk).
- Niech ten mężczyzna kupi mleko (Let this man shall buy milk).

There will be no difference between the above sentences for the detection algorithm. While in case of the conditional mood, the algorithm shall consider another form of the verb (compliant as regards the person, the number and the gender):

- Mężczyzna kupiłby mleko (The man would buy milk).

If a modal verb is detected, the algorithm analyses the sentence from the modal verb point of view. The implementation treats all verbs requiring presence of other verb in infinitive as modal verbs. The modal verb is not

stored after extraction. It is only used to indicate that the verb in infinitive form should be extracted from the sentence.

6. EVALUATION

The set of several dozen sample sentences was prepared as part of the test phase of the discussed algorithm. Detection results were manually verified for those sentences. In case of simple sentences comprising single-meaning words, all detections were performed correctly.
To increase clarity of the presented information, the results are displayed in form of a graph. The figure 3 shows a sample graph developed for the dictionary entry grać (play). The graph was generated automatically based on several sentences including the verb grać.

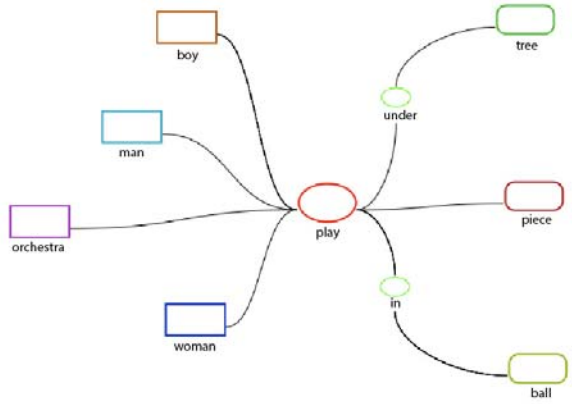

Figure. 3 Sample graph generated by the application.

The dictionary entry (verb) is displayed in the form of an ellipse in the centre. Nouns are presented in form of rectangles. Those with sharp corners are subjects of activities, while those with fillet corners – objects or tools.
However, there are many problems related to extraction of relations between words of the text. They result from ambiguity of words and functioning of the Inflection dictionary of Polish language, which analyses forms for all meanings of a given word in such a case.
There are words of numerous meanings, the inflexion of which is complicated, because it depends on the meaning, e.g. inflexion of *admirał* (admiral of the navy) slightly differs from the *admirał* (butterfly species). As a result, two known records are returned for the entry *admirał*.
The same situation occurs, when words are different, but they have some common form. For instance, the verb *mieć* (to have, to be the owner) and maić (to decorate something with green branches, leaves or with flowers) in Polish have a common form *mają*. That form, in a sample sentence below: *Nauczyciele mają czas na poprawę do końca tygodnia.* (Teachers have some time to check the test by the end of the week), can be interpreted two ways. Hence the graph includes two forms (as shown in the figure 4).

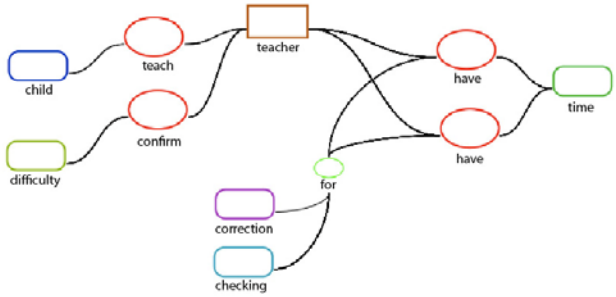

Figure.4 Sample graph showing incorrect determination.

Another important problem concerns situation, when different parts of speech are returned for a given word. It is extremely difficult task to determine, which one is correct. For example, the word jak returns known form of an adverb (which is much more frequent) as well as a noun (animal species). Thus we are able to determine many sentences with jak as a potential subject.
As it can be seen in the above examples, fully automatic generation of the entry description results in some senseless or totally incorrect matches, which are caused, among other things, by incorrect matching of the meaning or inflexion group during word detection.

7. SUMARY AND CONCLUSIONS

Authors of this paper proposed the idea of a visual dictionary, which, according to the general classification of dictionaries provided in the article [8] can be categorized as a nest-type dictionary. This type is specific, e.g. for synonym dictionaries and facilitates illustration of formal relations between entries.
The developed dictionary automatically generates visual description of each entry (in form of a graph), thanks to which it increases its advantage over paper dictionaries and even over digitalised versions of paper dictionaries. Using an appropriate internet robot, i.e. so called web crawler, capable to collect only simple sentences from the Internet, dictionary entries can be expanded as consecutive web pages are browsed.
Operation of the algorithm was evaluated using several dozen selected sample sentences. In case of simple

sentences comprising single-meaning words, all detections were performed correctly. Regrettably, in other cases, fully automatic generation of dictionary entry descriptions results in partially incorrect detection. Currently, the only solution of that problem is to manually verify stored entry descriptions. To extend the dictionary discussed in this paper, a thoroughly prepared corpus including simple sentences and the hardware capable to process huge amount of information would be required.

It is possible to improve the applied algorithm by manual development of the dictionary of verbs including information on the cases individual verbs are associated with, which would significantly increase word matching correctness. Moreover, it is still considerable challenge to analyse more complex sentences, in particular compound clauses.

Other dictionary extension options may take into account frequency of verb use and their individual attributes as well as prevalence of relevant combinations, which would have effect on selection of a given combination depending on the type of the statement.

References

1. Bień J. S., Linde-Usiekniewicz J., „Elektroniczne słowniki języka polskiego", Postscriptum (23-24), pp 17-25, 1998

2. Kluza K., Gatkowska I., „Przegląd polskojęzycznych słowników internetowych", 2011 (artykuł zgłoszony do publikacji w monografii)

3. Pisarek P., „Słownik fleksyjny", Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu, red. W. Lubaszewski, AGH Uczelniane Wydawnictwa Naukowo-Dydaktyczne, pp. 37-68, Kraków 2009

4. „Słownik fleksyjny języka polskiego", praca zbior., red. W. Lubaszewski, Grupa Lingwistyki Komputerowej, Kraków 2001

5. Szpakowicz S., „Formalny opis składniowy zdań polskich", Wydawnictwa Uniwersytetu Warszawskiego, wyd. 2, Warszawa 1986

6. Świdziński M., „Gramatyka formalna języka polskiego", Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 1992

7. Vetulani Z., Martinek J., Obrębski T., Vetulani G., „Dictionary based methods and tools for language engineering", Wydawnictwo UAM, Poznań 1998

8. Żmigrodzki P., „Wprowadzanie do leksykografii polskiej", Wydawnictwo Uniwersytetu Śląskiego, Katowice 2009