Morphonary: A Slovak Language Dictionary

Stanislav Krajči, Michal Mati, and Róbert Novotný

Institute of Computer Science, Faculty of Science,
Pavol Jozef Šafárik University in Košice,
Jesenná 5, 040 01 Košice, Slovakia
Name.Surname@upjs.sk

Abstract. We propose an algorithm for determining all flexions of given word for Slovak languages based on the principle of searching for similar words or "patterns". Furthermore, an analysis and design of tool encompassing these algorithms is provided.

1 Introduction and motivation

While processing new documents in a vector model that includes more or less sophisticated statistics of term occurences in a document it is useful to group distinct forms of one word and to choose a "suitable" representant of respective group. Usually a stem or a base form is taken as a representant. This process is however completely different for the two languages we take into account – English and Slovak.

In English we come across with a simple word flexion only which makes a search for a common base form easier. This is more clear when we notice relatively easily definable rules for deriving base form of the word. Of course there are some exceptions (i.e. irregular verbs) but of minor relevance as they are not numerous. The most common method is a stemming using the Porter's algorithm [1] based on a suffix removal (i.e. -ed, -ing or -s). Note that the number of considered suffices is small.

Slovak however is a highly inflective language with words having tens of various forms created by hundreds of means. Therefore it is enormously time-consuming to construct and apply some rules for a word flexion. This process could indeed consume much more time than a searching for all forms of all words manually. Hence it is more sensible to consider the second mentioned possibility when dealing with Slovak language.

Our aim in the first phase is to construct an electronical list (of base forms) from the Dictionary of the Slovak Language [2] broadened by words from the Dictionary of Foreign Terms that are commonly used but are not contained in SSJ. The second phase will encompass a semi-automatic extension of such word list by their various forms. Such dictionary would be able to find all forms of a search term, which would enrich a resulting set of documents (i.e. a search for "auto" would also return "áut" (genitive plural) as a result which would otherwise remain unnoticed by a simple search).

2 Description of Methods

As we have mentioned abundant experience in searching for forms of a word in English is unusable when processing the Slovak language. We use a very simple method of deriving forms of a word from its base form. The method is based on a trivial observation that flexion itself is realised by changing suffixes (not in linguistic sense) of a word. This means that if we try to find declinations of a substantive X, it is sufficient to find a "pattern" substantive Y (with the same gender) having its declinations already added into the database. It is necessery to find such Y that has the longest possible common suffix with X. We denote such suffix as Z. In this case it is highly probable that X can be declined in the same way as Y e. g. having the same suffixes for the respective grammar cases.

Let $X = X' + Z$ and $Y = Y' + Z$ (symbol + denotes string concatenation). If some case of Y is $Y' + K$ then expectably the same case of X is $X' + K$.

Example 1. Let $X =$ "migréna", $X' =$ "mig", $Y =$ "aréna", $Y' =$ "a". Note the different word bases. Now let's assume we know the singular instrumental of Y is "arénami". In the previous denotation: $Y' + K =$ "arénami", hence $K =$ "rénami". So the expected singular instrumental of the word $X =$ "migréna" would be $X' + K =$ "mig" + "rénami" = "migrénami".

3 Design and Implementation

We have designed and partially implemented a tool which carries out proposed methods. This tool *Morphonary* / *Tvaroslovník* is used in the project NAZOU[1].

It comes into use when it is necessary to find a base form of a word or to find all forms of a word – typically for expanding a query. It is designed to integrable with each tool involving usage of Slovak language especially dealing with words of unpredictable grammatical form.

Globally, the tool comprises of two logical units: the first provides components and user interface for the aforementioned semiautomatical construction of all forms of a given word. The implementation fully corresponds with the description of used methods. The second integration unit contains the interfaces which enable other tools to access the dictionary transparently. Both components are backed by a relational database.

The *Semiautomatical Data Input and Correction User Interface* has been implemented as a web application. Its main purpose is to allow users perform the correction and maintenance of word data and to help in exercising the flexion process of the system.

The *Morphonary* encapsulates the inflecting algorithm. The input is generally a word for which we need to retrieve flexions. Its output can be used in two ways:

[1] NAZOU = "Tools for acquisition, organization and maintenance of knowledge in an environment of heterogeneous information resources", homepage: `http://nazou.fiit.stuba.sk/`.

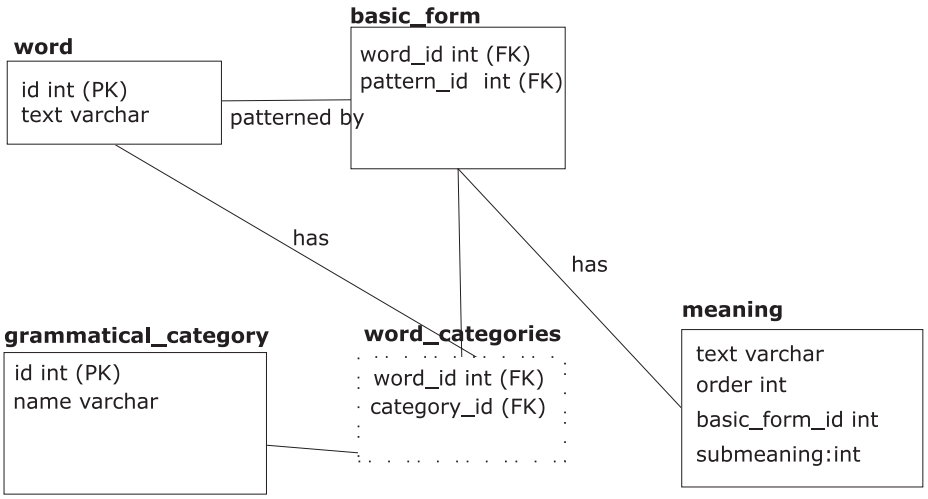

Fig. 1. Entity-relationship model of the fundamental tables.

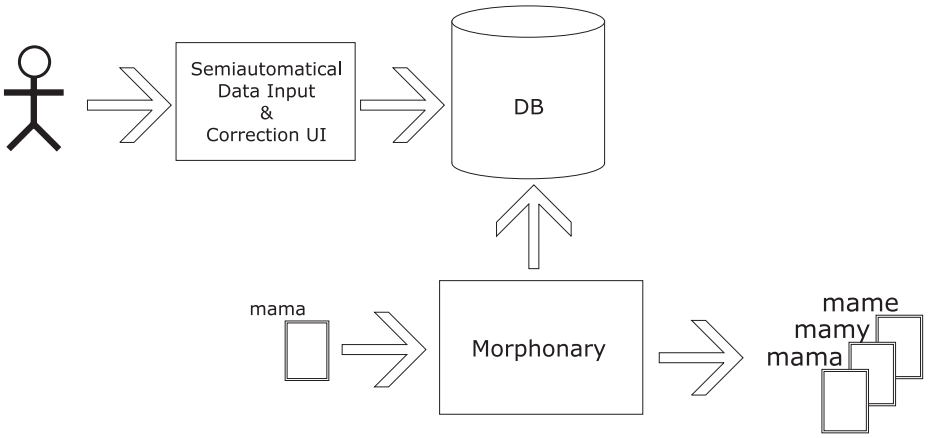

Fig. 2. Schema of the Morphonary Tool.

either as a proposal for user to the Correction UI, which is then adjusted by the user, or as the set of flexions, which can be used by another tools.

This component also contains the metrics for finding the best patterns according to various parameters, thus solving the problem of too many patterns found. Current design of the algorithm requires that user provides at least the lexical category, and in the case of substantives, at least the gender. This minimizes the amount of erratic proposals.

The *database backend* is currently implemented in the IBM DB2 relational database. Database model of the fundamental tables (we left out the auxiliary tables for brevity) is show in the figure 1.

4 Future work

The future work will be focused on experimenting with various metrices which should eliminate situations in which the user is overwhelmed by large number of incorrect proposals.

Furthermore, we plan to reimplement the Data Input User Interface to allow its deployment and management in the multi-user environment, thus allowing to use external human resources in the dictionary data maintenance.

Acknowledgements

This work was partially supported by the Slovak State Programme of Research and Development "Establishing of Information Society" under the contract No. 1025/04.

References

1. Porter, M. F.: An algorithm for suffix stripping, Program, 14(3) pp 130-137, 1980.
2. Slovník slovenského jazyka, Volumes I.-VI., Vydavatel'stvo SAV, Bratislava, 1959-1968
3. Páleš, E.: Sapfo, parafrázovač slovenčiny, Veda, Bratislava, 1994. ISBN 80-224-0109-9.