Electronic Dictionaries: For Both Humans and Computers

Igor A. Bolshakov, Alexander F. Gelbukh, and Sofia N. Galicia-Haro

Natural Language Laboratory,
Center for Computing Research (CIC), National Polytechnic Institute (IPN),
Av. Juan de Dios Bátiz, CP 07738, Zacatenco, Mexico City, Mexico
{igor, gelbukh, sofia}@pollux.cic.ipn.mx

Abstract. The modern electronic dictionaries of natural languages should be universal. In the linguistic aspects, they should be a multi-linked database similar in their contents to the combinatorial dictionary by I. Mel'čuk, but with more stress on thesaurical links and word combinations. In interface aspects, they should have their data accessible to a text processing software, a human user and a lexicographer.

1 Introduction

During several decades, two different types of dictionaries of natural languages existed in parallel. In the paper form, they were oriented to various readers, in the electronic form, to needs of automatic text processing.

Recently electronic dictionaries have also appeared repeating a paper form and directly oriented to a human. All limitations on the size of dictionaries and on the complexity of demos on the screen were eliminated, with the tendency to minimize the role of the paper. Some computer scientists consider this as an ultimate solution of the problem of electronic dictionaries, but the situation is not so optimistic. The ex-paper dictionaries, even academically complete, do not contain all information necessary for text processing, and no automatic procedure can derive this information from human-oriented texts.

We argue for a universal dictionary, similar in its contents to [1], but with more stress on word combinations and thesaurical links. Three groups of possible use of the universal dictionary and some requirements oriented to various applications are described.

2 Some Deficiencies of Human-oriented Dictionaries

Trying to use the contents of two big electronic dictionaries of Spanish [2,3] for automatic processing of texts, the authors found a lack of information of graphical, morphological and syntactical nature. Indeed, no automata can calculate what lexemes in the pairs *lunes* vs. *mes* are invariable. All dictionaries give the labels of transitive verbs, adjoining pronominal clitics in accusative case. Meanwhile, for the group of dative verbs the number of agglutinated clitics that can be up to two, and without label of dativity to process such forms is impossible.

V. Matoušek et al. (Eds.): TSD '99, LNAI 1692, pp. 365–368, 1999.

In English, morphological peculiarities (nonstandard endings of plural for nouns like *phenomenon*, nonstandard paradigms for such verbs as *do, see, go*, etc.) are given in Merriam-Webster and other dictionaries. For Russian, a formal representation of its sophisticated morphology was given 20 years ago in [4] and immediately adopted by numerous software developers.

However, attempts to find in academic dictionaries combinatorial properties of words are usually in vain. The information about valences of nouns, adjectives, and especially verbs is scarse, even for English. Therefore, we cannot know how to express in Spanish or Russian combinations like *to pay attention*.

3 Some Deficiencies of Computer-oriented Dictionaries

The main problem with computer-oriented dictionaries is the same: each of them contains only specific kind of information, so that several dictionaries are necessary to process the information on different language levels. On the first glance, computer dictionaries can be easily combined with each other, but it is so.

First, the sets of words in various dictionaries are different. Except for a small kernel, dictionaries tend to differ very significantly in their coverage. Second, combining dictionaries is not at all straightforward. The result is consistent only if the corresponding senses of the homonymous words are combined correctly. However, both the number and the sequence of the senses in different dictionaries are different, and there is no way to automatically recognize mutual correspondences. Computer dictionaries lack in human-oriented remarks, grammar reference, tables of abbreviations, etc.

4 Idea of the Universal Dictionary

Hence, the problems of the existing dictionaries are caused, apart from their natural incompleteness, by disruptive information spread across various sources. One needs to look up (and probably search for) many of them to see all about a particular word. Thus, our main idea is rather trivial:

- *A computer dictionary must present all the information about each word and the language as a whole.*
- *It must present all the possible ways of accessing and searching this information.*

We mean that the information should be accessible to both human users and other programs. A dictionary is so large and expensive database, that it is better to maintain, keep, and use its common version for all applications. It should be presented uniformly and be available in an integrated environment, such as a common browser (for users) or Application Program Interface (for programs).

The universal dictionary must also give all the available grammatical information, with all necessary cross-references. Since grammatical information may have a form of algorithms, the dictionary should not only show texts and describe algorithms, but also provide programs realizing them, e.g., various tools

of checking and parsing. The new dictionary would not be a mechanical combination of different sources, though it is hardly possible to organize right now a great project on creation. All available sources should be merged by a program parsing various formats and compileing all data to a consistent whole.

The important problem is to avoid repetitions of in the common entries. Well-formalized information like morphological can be easily uniformed and merged. However, it is not the case for the explanations. At the same time, the merge should involve minimum of manual work. In any case, the number of possible updates should be great.

As the idea of the universal dictionary becomes more popular, it will be possible some standards, for unification of the formats of the sections. This will give lead to better procedures for further merges. If the source dictionaries cover different domains, it is enough to mechanically combine them, adding special marks to combined parts.

5 Contents of the Universal Dictionary

The universal dictionary should ideally contain the following types of information:

- Orthographic form of the keyword, including options and standard abbreviations
- Pronunciation including options
- Phonemic features, especially syllabic structure
- Morphological features: part of speech, inclination or conjugation class, etc.
- Syntactic features
- Explanations, maximally structured and consistent, including allusion and style features
- Semantic references, at least to synonyms and ideally to a thesaurs or a semantic network
- Combinatorial features, in style of a full combinatory dictionary and/or the dictionary of word combinations
- Equivalents in other languages
- Examples of usage

The necessity in the dictionary of the combinatorial information should be especially emphasized, since it can be currently found only in special dictionaries [5]. To compile lists of co-occrring words is much easier as compared with listing the lexical functions. Meanwhile, together with a thesaurus this facility proves to be very useful for both users (text compilation) and programs (syntactical analysis and disambiguation).

6 Needs of the End User

For a common user, the dictionary should provide a browser giving necessary data from the linguistic database in a convenient form. The interface developers

should take into account, that: there is no need to save the space in the dictionary; the colors, fonts, etc., can be wider used; a nested hierarchy of paragraphs is much better for an entry than a single paragraph; the examples can be used more intensively. The data can be customized on the screen, with removal of unwanted parts. The data should be sorted by various categories. The request can combine logical means, such as AND, OR, NOT operators.

The dictionary should also give access to word-formation, agreement within word combinations, determining the syntactical structure of a phrase, language learning, etc.

7 Needs of the Text Processing System

For text processing software, the dictionary should have a library of procedures permitting to service any separate step of language processing or all of them. All the data are accessible from other programs in a formal way. The inner representation should not be just strings from a paper source, but members of well-defined sets. The dictionary should service various other programs, from spell-checker to text translators. There is no need to wait when these utilities are brought to perfection; they should be available right now.

8 Needs of the Lexicographer

The universal dictionary should be the environment to elaborate new data for this or other dictionaries. It should provide a way to modify the information, make temporal notes, etc. It should contain a specialized language to create the private programs for investigating lexical data. Pieces of data should have labels of its completion. Those without the labels are incomplete or accessible only to privileged users, for further elaboration.

References

1. Mel'čuk, I. A., Zholkovsky, A. K.: Explanatory Combinatorial Dictionary. In: Even, M.W. (ed.): Relational Models of the Lexicon. Cambridge University Press, (1988) 41–74.
2. Diccionario del Español contemporaneo. Grupo ANAYA, `http://www.anaya.es`.
3. Diccionario de la lengua Española. Real Académia Española, Edición en CD-ROM (1996).
4. Zaliznyak, A.A.: Grammaticheskij Slovar' Russkogo Yazyka (Russian Grammar Dictionary). Russkij Yazyk, Moscow (1974).
5. Bolshakov, I.A.: Multifunction Thesaurus for Russian Word Processing. In: Proc. 4^{th} Conf. on ANLP. Stuttgart (1994) 200–202.