recibido 18-05-2007; aceptado 22-06-2007

Portal da Língua Portuguesa

Maarten Janssen
Instittúto de Linguística Teórica e Computacional (ILTEC)
Rua Conde de Redondo 74-5, Lisboa, Portugal
maarten@iltec.pt

Resumen: El objetivo del proyecto *Portal da Língua Portuguesa* es construir, con un doble objetivo, un juego de recursos léxicos. En primer lugar, estos recursos sirven como fuente de información para una página web sobre la lengua portuguesa para el público en general. En segundo lugar, son un repositorio de información léxica para la investigación lingüística. El dibujo de la base de datos es modular y relacional, y se hizo de modo que proporcione soluciones estructurales para problemas léxicos, como son los de la homonimia, variación ortográfica, etc.
Palabras clave: Base de datos léxica, morfología, fonética.

Abstract: The goal of the *Portal da Língua Portuguesa* project is to construe a set of lexical resources with a double objective. On the one hand, the resources serve as the content source for a web site about the Portuguese language, aimed at the general public. On the other hand, the resources are built to serve as an open source repository of lexical information for linguistic research. The design of the database is modular and relational, and is set-up in such a way that it provides structural solutions for lexical difficulties like homonymy, orthographic variation, etc.
Keywords: Lexical database, morphology, phonetics

1 *Project Description*

The *Portal da Língua Portuguesa* (henceforth Portal) is a free, large scale online resources on the Portuguese language, currently under development at the ILTEC institute in Lisbon, Portugal. It has a primary focus on lexical information, and is designed for the general language user. Although the Portal is the visible outlet of the Portal project, the goal of the project itself is moreover to create a set of lexical resources which, apart from their online availability, will serve as open source data for linguistic research. The project started from lexical database called *MorDebe*, which primarily concerns inflectional morphology. But the database is currently being transformed into an Open Source Lexical Information Network (OSLIN), which contains a much wider, open-ended range of lexical information. Additional types of lexical information currently under development are inherent inflections, pronunciation, and syllabification.

The Portal project itself is internally supported by the ILTEC institute, and has no strict delimitation. Work on the MorDebe database was started mid 2004, and the web site was launched in November 2006. The web site is intended to continue for an undetermined amount of time. The project has two full-time FCT-funded scholars assigned to it for a period of 3 years, starting from September 2006. The project is enforced by satellite projects, which deal with specific parts of the database. A two-year project on the improvement and exploration of the derivational data in OSLIN will start in October 2007, and run for two years.

2 *OSLIN Design*

2.1 Main database

The main database of OSLIN (MorDebe) consists of a simple two-table structure, one table with lemmas, the other with the related word-forms. The lemma list consists of two parts – on the one hand, it contains the lemmas from the two major Portuguese dictionaries, and on the other hand, it contains words with a significant frequency in newspapers. In both parts of the database, a strict lexicographic control is kept over the data, with a significant amount of human intervention, using computer-aided methods. The total number of lemmas at this moment is around 130k, with constant additions being made, and well over 1,5M word-forms.

Although the MorDebe database was set-up for Portuguese, its design is largely language independent. The set of word classes and inflectional forms is determined in a separate database, and can easily be modified to accommodate languages with rich nominal inflection, or with other fundamental word classes.

2.2 Inherent Inflection

In the database, inherent inflection (Janssen, 2005) are modelled in terms of relations between lemmas, using relations similar to those in the Meaning-Text Theory (Mel'cuk, 1993) called inflectional functions. With these inflectional functions, verbs are related to their deverbal nouns (s0v), adjectives to their synthetic superlative (sup), etc. The inherent inflection database is still under construction, and contains currently over 20.000 derivational forms. It is planned to feature the complete set of all dictionarized inherent inflections within the scope of a year.

There are two types of relations that are modelled in a way similar to inherent inflections, but are of a different nature. The first is a separate database of gentiles: all nouns and adjectives indicating people or objects from a specific space or region are relationally marked as such. The difference with inherent inflection is that toponyms are not lemmas, and are stored in a separate database of proper names. The complete set of all over 3000 dictionarized gentiles has been modelled in this fashion.

The second special type of 'inflectional function' is the relation between orthographic variants. Orthographic variation is traditionally seen as an intra-word phenomenon. But the explicit modelling of inflectional paradigms makes it necessary to keep the different variants apart and interrelate them with a relation (Janssen, 2006).

2.3 Web Site Design

The web-site of the Portal provides (or will provide) five different types of information: not only the lexical information from the MorDebe database, but also information on legislation, a dictionary of linguistic terms, a repository of online resources on Portuguese other than the Portal itself, and a collection of easy texts concerning the Portuguese language. With the current content, the web site already attracts some 1000 visitors each day, mainly language professionals such as translators and writers, and that number is steadily rising.

The use of the MorDebe data in an online service for the general public provides an excellent additional motivation for the creation of the lexical resources, and even opens up the possibility of commercial sponsoring.

2.4 Modular Design

The design of the OSLIN database is fully modular: each additional type of information is modelled in a separate database, linked to one of the existing tables, currently either the word-forms or the lemmas. This design makes it easy to extend the database with additional types of information. The main resource currently under development is a database of IPA transcriptions for all lemmas in the database, but various other types of information are under investigation. At this time, there are no plans to add semantic entities, merely due to lack of resources, not because the framework does not allow it. Ideally, the framework would be extended to other languages besides Portuguese in the near future. Using the same set-up for various language would not only allow reusing the existing tools, but also make it possible create cross-linguistic relations.

Bibliografía

Janssen, Maarten. 2005. "Between Inflection and Derivation: Paradigmatic Lexical Functions in Morphological Databases". En *East West Encounter: second international conference on Meaning - Text Theory*, Moscow, Russia.

Janssen, Maarten. 2006. "Orthographic Variation in Lexical Databases". En *Proceedings of EURALEX 2005*, Turin, Italy.

Mel'cuk, Igor A. 1993. The Future of the Lexicon in Linguistic Description. En Ik-Wan Lee (ed.*) Linguistics in the Morning Calm 3: Selected papers from SICOL-1992*. Korea: Seoul.