MorDebe-Admin
A Lexicon Management System

Sílvia Barbosa, José Pedro Ferreira, Maarten Janssen
ILTEC, Lisboa – IULA, Barcelona

1. Introduction

MorDebe-Admin (henceforth MA) is a lexicon management system developed at the ILTEC institute in Lisbon, Portugal. It was developed as the dedicated administration environment for the maintenance of the data of a large-scale lexical database system called *MorDebe*[1] (Janssen, 2005a). The set-up of the database, as well as its maintenance system, is largely language independent, although so far the system is only under development for Portuguese.
The MorDebe database for Portuguese at this moment contains around 130.000 entries, and 1.3 million word-forms. It is a lexicographically maintained lexicon that is more and more frequently taken as a reference work for Portuguese.
This article describes the MA system from the design perspective, but more prominently from the perspective of the lexicographer working with it. Before describing MorDebe-Admin itself, the next section presents a brief overview of the MorDebe database it administrates, as well as the web page for consulting the database.

1.1. MorDebe Design and Portal

MorDebe consists of a number of individual databases, each of which describes a specific aspect of the lexicon. There is a database that contains the lexical items of the language, with their word class and citation form. Another database describes all the inflected forms of each of these lexical items. Yet other databases provide the pronunciation and syllabification of the entries, and the (derivational) relations between them. All these individual databases are linked into a single relational database by means of cross-references. Additional databases are still constantly being added to MorDebe, describing further aspects of the lexicon. For the moment, only formal aspects of words are described, leaving out all information that is meaning or etymology related.
The content of the MorDebe database can be consulted online by the general public on a website called *Portal da Língua Portuguesa* (portaldalinguaportuguesa.org - henceforth Portal). The Portal is the main interface to the MorDebe database, and is oriented towards the common language user, with a special emphasis on the undergraduate language learner. However, the site is intended for the scientific community as well. The Portal not only presents the lexical data from the MorDebe database, but also a number of other useful resources concerning the Portuguese language, including the text of the *acordos ortográficos* (official spelling), and a dictionary of linguistic terms. The Portal is a well-visited web site with currently around 2.000 visitors per day, and attracting more and more visitors all the time.

2. MorDebe-Admin

[1] In principle, MorDebe is only a part of the larger Open Source Lexical Information Network (OSLIN – Janssen, 2005b), and MorDebe-Admin is used to edit all parts of OSLIN. For the sake of clarity, we will refer to the entire lexical database here as MorDebe.

MorDebe-Admin provides an integrated collection of tools dedicated to the maintenance of the individual databases of MorDebe. It also features some tools for the expansion of the lexicon, such as several corpus tools and a built-in neologism tracking system. This chapter describes the implementation of MA, the editing and checks, and some of the specific tools.
In a database the size of a full-scale lexicon it is inevitable that minor errors slip in during the manual and semi-automatic processing, especially in a database with a modular design such as MorDebe. MA is designed to prevent such errors in two ways. Firstly, it takes care of all the editing centrally and it checks and updates all the related databases on every update, insert, or delete in any particular database. Secondly, it provides a number of checks that are run on a regular basis to verify whether no inconsistencies have arisen despite the use of the central editing.

2.1. Background

Many of the existing Dictionary Writing System (DWS) are in a sense dedicated text editors, mostly based on the XML file format. The core of such system is a model (called the DTD) in which you specify the overall structure of the dictionary entry – which fields it has, and which fields are obligatory or optional. Although some fields can only be filled in with a limited set of values, most fields will just be text boxes where the lexicographer will fill in the etymology, the meaning definition, the example sentences, etc. Advanced DWS systems such as TshwanaLex (de Schryver & de Pauw, 2007) come bundled with rich sets of features that help the lexicographer to define and look for words such as part-of-speech taggers, concordancers, etc. Furthermore, these many DWS are structured in such a way that various dictionaries can share a common set of data on pronunciation, word class, etc. But the core function of most DWS is to make the process of creating the graphical end product (the dictionary itself) easier, while allowing maximum freedom to the lexicographer.
The idea behind MorDebe - and hence behind MA - is quite different. MorDebe has very few free-text fields, except for the notes, the citation forms, and the orthography of the entries in the other databases. This is largely because MorDebe-Admin is not a DWS – most significantly, MorDebe does not contain information about meaning, nor about etymology. MA is more comparable with the VisDic editor for WordNet (Horák & Smrž, 2003), which edits relations between items.
As said in the introduction, MorDebe consists of a set of related databases. MA provides a way to manage these different databases in an easy way, while maintaining the consistency of the entire relational database. MorDebe provides some windows that collect information from different databases, as well as dedicated editors for each of the individual databases. As more databases are being added, MA is also being extended with additional modules for the maintenance of these additional databases.

2.2. General overview

MA consists fundamentally of four modules: *add/edit*, *checks*, *resources*, and *varia*. *Add/Edit* is the main way of adding new lexical items, and editing existing ones. All entries in MorDebe are dealt with via this module. The information about each lexical entry is physically distributed over different databases, but the editing window collects the relevant information from the various databases and presents them in a single form (see figure 1). In the main edit window, it is possible to edit the following information: the dialectal restriction of the lexical entry, the syllable division, the phonetic transcription(s) of the lexical entries, the notes, all inflected forms

associated with the lexical entry, the sources in which the words can be found, and the derivational relations. Apart from these resources that can be edited directly from the edit window of a lexical entry, the edit window also provides links to the specific editors some related database (such as the databases of gentiles and loanword) when these databases have an entry for that word. When adding a new lexical entry, MA automatically provides the inflectional paradigm, the pronunciation, etc. (as described in the next chapter). After each add or edit, there is a direct link to visualize the result on the Portal pages.

Figure 1. The main edit window of MorDebe-Admin

In the *checks* section, it is possible to run several consistency checks over the databases, to find potential errors. These checks provide lists of words that have characteristics that should not occur - the checks do not correct anything automatically, but merely present links to the edit window of the suspicious entries and relations, so that the lexicographer can easily correct the errors. There is for instance a check that verifies whether all lexical entries have at least one inflected form, and whether there are citation forms that start or end with a space. The consistency checks are described in more detail in section 3.3.
In the *resources* section, one can edit the information of the various smaller dictionaries in MorDebe (see section 3.4), as well as edit some databases providing non-lexical information that is presented on the *Portal da Língua Portuguesa*. In the section of the smaller dictionaries, there are specific editors for all the individual lexical resources in MorDebe: the loanword dictionary, the gentiles dictionary, the derivational databases (deverbal nouns, female nouns, etc.), and the pronunciation database. For all of these, it is possible to generate an alphabetic list of all the entries, to search the database, and to edit all the individual entries.
The *varia* section consists of a collection of different tools, such as the feedback system (see 3.6) and the interface with NeoTrack (see 3.5), as well as some links to useful online resources.

3. System Details

3.1. System Requirements and Implementation

MorDebe-Admin is a completely Internet based system. That is to say, none of the editing, adding, or checking is done using programs installed locally on a personal computer or using data stored on a local hard disk, but everything is done via the Internet. This means that MA allows different people to work with the same data at the same time, even if they are in completely different locations. For a language like Portuguese, which is the official language in countries across several continents, this is a very useful feature. Since the Internet access means that anybody can in principle reach the system, access is restricted by a secure login.
All changes made in the system are recorded in a log-file, together with the date and editor. Although the system does not have a so-called *roll-back feature* allowing you to undo the last changes, the log-file makes it possible to check when things where last edited.
MA is an open source web based software package, written in a combination of the PHP and Perl programming languages and the MySQL database system. It does not rely on any commercial packages, and can easily be installed on any UNIX based server. The system consists of a number of individual scripts, each of which takes care of a specific part of the overall maintenance and editing.

3.2. Inserting New Words

When a new lexical item is added to MorDebe, it has to be added with all its inflectional forms. MA contains a built-in regular inflection mechanism that generates all the inflected forms automatically to reduce work and avoid errors. The system also produces the syllable division and the pronunciation (IPA transcription) for the lexical entry automatically. Since there could be irregular or unpredictable forms in the inflectional paradigm, the syllabification, or the pronunciation, none of this information is added fully automatically. Instead, the suggested form are merely pre-filled into the relevant fields of the add form, and should always be checked by the lexicographer. In cases where there are known problems with the automatically generated information, MA even presents the problematic forms in red. An example of a problematic form is the plural form of hyphenated compounds in Portuguese, which can inflect on the left (*palavras-chave* = "words-key", keywords), on the right (*guarda-chuvas* = "guard-rains", umbrellas), on both parts (*homens-fortes* = "men-strong", strongmen), or nowhere (*fora-da-lei* = "out-of-law", outlaws). Which part inflects cannot be predicted on the basis of the orthography alone, and should always be decided by the lexicographer.
MA also contains a lightweight morphological analyser, which identifies words as potential gentiles, deverbal nouns, superlatives, etc. When a word is detected as a likely derivative or gentile, an entry for the relevant database is suggested by the system.

3.3. Consistency Checks

MorDebe-Admin provides a number of consistency checks to verify potential errors in the databases. These checks are needed because in a large relation-based system like MorDebe, it is inevitable that errors occur, despite careful work and the use of the constraints put out by the use of MA. For instance, when adding a lexical entry it is too easy to overlook possible spaces before or after the word. Therefore, the system provides an easy way to find all words that start or end with a space, or contain unexpected characters, and gives a list of all of them. Each item on the list contains a link to the edit window for the related entry in the relevant database.
MA takes care of deleting all the inflected forms of a lexical entry when the entry is deleted. Since there is no fundamental restriction on entries without inflected forms, is possible that somehow there are lexical entries without inflected forms, or inflected forms without an associated lexical entry. Therefore, there are also checks that list all the lexical entries and inflected forms that lack their counterpart.
Other checks in MA are the following. A check for potential duplicate entries, that is to say, entries with identical citation form, morphosyntactic class and subclass. Consistency checks on the derivational relation, verifying for instance whether all deverbal noun relations are indeed relations between a verb and a noun. A check to see whether the citation form is identical to the expected inflected form for that class, e.g. whether the citation form of the adjective is the same as the masculine singular form. And last but not least, a list of all the latest additions to verify whether despite all the checks and restrictions no errors have slipped through. No corrections are made fully automatically - a lexicographer always verifies all suggested corrections.

3.4. Editing Mini-Dictionaries

Many of the data on the Portal are organized as mini-dictionaries. For instance, there is a dictionary of gentiles and toponyms, a dictionary of deverbal nouns, and a dictionary of loanwords. For the maintenance of these mini-dictionaries, MA integrates specific modules for editing and adding items. For instance, there is a module for editing the loanword dictionary. The module allows you to browse or search all the entries of the dictionary, to view the individual

entries, and to edit the content of each entry. Editing or adding entries of the dictionary automatically modifies the entries of the underlying databases. As said before, new databases and mini-dictionaries are still being added to MorDebe / Portal. The modular design of MA allows for the easy addition of complementary modules for the maintenance of any additional dictionaries.
Some of mini-dictionaries in MorDebe, such as for instance the database of gentiles, have a complex internal structure – the gentiles dictionary involves both a database of proper names, and a database linking the proper names to lexical entries. Furthermore, the choice of which gentiles to include involves strict criteria, given the enormous amount of toponyms. For these databases, MA only allows a restricted set of members to directly edit the dictionary. Other members can only add entries to a temporary database, in order to better control the quality.

3.5. NeoTrack

Language change implies that lexicographic work is constantly getting outdated - new words appear, old words change or disappear. The fact that a large amount of text is available in digital form nowadays makes it possible to automate the process of detecting neologisms. However, the introduction of neologisms should still be under manual control to avoid the insertion of typographic errors and occasionalisms.
For the treatment of neologisms, MA integrates a neologism-tracking tool called *NeoTrack* (Janssen, *forthcoming*), that is capable of identifying words that are not in MorDebe in the online versions of newspapers (or other HTML documents). NeoTrack uses a list of known words (the exclusion list) to determine which words are neologistic, instead of the frequency-based criterion used in some other neologism tools such as NeoloSearch (Janicijevic & Walker, 1997). The NeoTrack tool is integrated with MorDebe in the sense that the exclusion list used is exactly the list of word-forms in MorDebe, and that new words found with NeoTrack can be fed back into MorDebe.
NeoTrack automatically creates a list of unknown words, or neologism candidates, encountered in the source text. It then helps to classify these candidates in errors, neologisms, and words that should have been in the exclusion list (gaps). The gaps are added to a temporary database, where they are listed for verification and rapid introduction in MorDebe.

3.6. Feedback

The Portal encourages its visitors to provide feedback for any potential errors they encounter. MA lists all comments together with their date, the page the comment originated from, and the word or entry the comment is about. The list makes it easy to treat the comments – it takes care of sending the response emails as well as storing all answers, it includes a number of template answers to frequent issues, it provides an easy way to get rid of spam, and it links to entries that are suggested as potential errors, making it easy to update the information.

3.7. Internal and External Notes

For a database the size of a full dictionary, it is important to be able to make comments with each item. Therefore, all databases in MorDebe come with a field for notes. These notes come in two types - internal and external notes. The internal notes are any type of remark that might be useful in the future: indications of problems to be looked at again, remarks on why something was changed, observations found in reference works, etc. These internal notes will only show up for the members of the team, to help in the maintenance of the database.
But many databases in the system also come with external notes - that is, notes that are visible to the end user of the *Portal*. Although the set-up of MorDebe is strictly relational, avoiding free-

text fields wherever possible, there is inevitably information that is important to display to the user, but for which there is not (yet) a structural treatment. For example, there are no usage labels in MorDebe at this moment. Therefore, the fact that the plural *anãos* for the Portuguese noun *anão* (dwarf) is infrequent with respect to the alternative *anões* is not currently modelled in the database, and put in an external note.

4. Conclusion

We hope to have demonstrated the general structure and use of the MorDebe-Admin system in this article. MA provides a flexible and user-friendly way of maintaining the distributed lexical resources of MorDebe. Although currently only in use for Portuguese, the system could easily be applied to other languages as well – all language specific tools, such as the automatic inflection tool, are centrally located and can be provided for all additional languages. Since the system has a very modular set-up, it is also easy to develop specific resources for the particularities of other languages.

Several areas of the lexicon are not yet treated by MorDebe or MorDebe-Admin. One of the main areas currently still under development is the treatment of Multi-Word Expressions (MWE). MWE are currently not really treated in MorDebe, and are currently only stored in a temporary database, which is not feature-rich enough to account for the complexity of MWEs, such as the amount of freedom to insert words they allow, whether and which parts inflect, which are the words that make up the MWE, etc. A better treatment of MWEs, as well as the treatment of some other aspects of the language, are still under development.

References

Horák, A. & P. Smrž. 2003. VisDic – Wordnet Browsing and Editing Tool*, In: Proceedings of the Second International WordNet Conference – GWC 2004*, Brno, Czech.

Janicijevic, Tatjana & Derek Walker. 1997. NeoloSearch: Automatic detection of neologisms in French Internet documents. *In: Proceedings of the Joint International Conference ACH-ALLC'97*. Kingston, Canada.

Janssen, Maarten. 2005a. Lexical vs. Dictionary Databases: design choices of the MorDebe system. *In: Papers in Computational Lexicography - COMPLEX 2005*. Budapest, Hungary.

Janssen, Maarten. 2005b. Open Source Lexical Information Network. *In: Third International Workshop on Generative Approaches to the Lexicon*. Geneva, Switzerland.

Janssen, Maarten. *forthcoming*. Orthographic Neologisms: selection criteria and semi-automatic detection. *Unpublished manuscript.*

Schryver, Gilles-Maurice de & Guy de Pauw. 2006. Dictionary Writing System (DWS) + Corpus Query Package (CQP): The Case of TshwaneLex. *Lexicos*, vol 17: 226-246.