Lexical vs. Dictionary Databases

Design Choices of the MorDebe System

MAARTEN JANSSEN
Instituto de Linguística Teórica e Computacional
Rua Conde de Redondo, 74-5 – Lisboa, Portugal
maarten@janssenweb.net

Many lexical databases are modelled simply as digital version of paper dictionaries. However, for many purposes the demands on a lexical database are different from those on a dictionary database. Therefore, the MorDebe database system deviates from the design of dictionary databases in a number of important ways. Firstly, it puts different restrictions on the inclusion of words due to its lesser restrictions in size. Secondly, it does not list only lemmas, but complete inflectional paradigms. And thirdly, lemma separation is form-based rather than meaning-based. This article discusses the advantages and problems of this different approach.

1. Introduction

Over the last few decades, a large amount of new lexical resources have arisen: machine readable dictionaries, lexical databases, full-form lexicons, morphological databases, semantic networks, dictionary databases, etc. Most of these lexical systems have been modelled after lexicographic sources. This paper discusses the design of a lexical database system called *MorDebe*, and why its design differs in important respects for the traditional set-up of dictionaries and dictionary databases.

The term *dictionary database* will be used in this article for a database whose primary function is the compilation of lexicographic products. This can either be simply a digital version of a paper dictionary, from which the printed version is generated (often called a machine readable dictionary), or it can be a complex system from which a wide range of monolingual and bilingual dictionaries are derived, as is the case with the Van Dale Lexicographic Information System (VLIS).

A lexical database, on the other hand, is a lexical resource system meant primarily for computational exploitation. This can be the use in a search engine providing human users with lexical information, but also the use in NLP applications, computer aided language-learning systems, computer aided linguistic research, etc. The lexical database system described in this article is called *MorDebe,* a system which aims explicitly at the use of a single set of lexical data in a wide range of applications, including both NLP systems and human consultation.

This article will discuss the main points in which the MorDebe database differs from dictionary databases. This comparison will be strictly from the perspective of the formal properties – with three main sources of difference: the amount and type of information stored for each lemma, the number of lemmas recorded, and the separation of lemmas. The discussion will focus not only on the motivations for these differences, but also on the resulting problems. Although a strict separation is not possible, the semantic properties of lexical databases are largely ignored in this article.

2. MorDebe Set-up

MorDebe is a lexical database system, whose set-up is largely language-independent, but whose content is currently purely Portuguese. In its current

version, MorDebe only specifies formal properties of words - the semantic component has not yet been developed. In the long run, MorDebe is intended to be integrated with the multilingual SIM*u*LLDA system (Janssen, 2002), providing formal, semantic, and cross-linguistic information.

The core of the MorDebe database consists of two related tables: the first is a table containing lemmas, defining for each lemma its citation form, its grammatical category, and when applicable, its compositional structure, and its terminological domain. The second is a table containing word-forms, specifying for each word-form its orthography, the lemma it belongs to, it inflectional form (number, gender, person, tense, aspect, etc.), and when available its syllabification and pronunciation.

The design of MorDebe aims at reusing the same set of data in a wide range of (linguistic) applications. To that end, the design is as much as possible theory and application independent. There are currently a number of ways in which the MorDebe database is used, including a part-of-speech tagger, and the analysis of derivational forms. Of these, two are particularly relevant for the current article:

MorDebe on-line consultation
The most direct use of the MorDebe database is its on-line consultation: there is a internet page that allows users to consult the MorDebe database, either via the lemma database, giving the stored information and the complete inflection, or via the word-form, giving the lemma it belongs to, as well as its inflectional form, the information about the lemma, etc. (more on this in section 3.2)

NeoTrack: neologism detection
The MorDebe database is used to generate the exclusion lexicon used in the semiautomatic detection of neologisms in on-line newspapers. The database is used bidirectionally: not only for the detection of neologism candidates, but new words encountered are also added to the database (more on this in section 4.1).

3. Lemma scope: full-form vs. lemma-only lexicons

Traditionally, dictionaries consist of (printed) lists of words, represented by their citation form, which the user can browse through. In some dictionaries, the key inflectional forms are represented along with the lemma of the entry, at least in the case of irregular inflection. But the dictionary is not commonly intended to be a complete source of inflectional information.

MorDebe, on the other hand, contains the full set of inflectional forms. This information is crucial to the lexical database, whereas it is largely irrelevant for traditional dictionary purposes. Other than what are usually called *full-form lexicons*, MorDebe does not merely list word-forms: all information is organised around lemmas. But for each lemma, its full inflection is provided.

3.1. Word-form driven database access

A full-form lexicon is clearly necessary for NLP use: the computer has to be told all inflected forms explicitly. But even for human consultation, there is a clear advantage of a full-form lexicon: database access is often most conveniently accessed by word-form, and not by lemma. Although it is common to look up words in dictionaries by their citation form, this is not always the most user-friendly solution. Especially for non-native users, it is sometimes hard to find words if the citation form is unknown – the word *mice* is hard to find if you do not know it is the irregular plural of *mouse*. And this is worse for prefixing language – for instance, it is hard for non-native users to find perfective verbs in Slavic languages. When trying to find the word *продублированный* in a the Oxford Russian-English dictionary, one has to know that it is a form of the verb *дублировать* (to duplicate), located at the other side of the dictionary.

In traditional dictionaries, this problem is often solved by putting irregular or hard-to-find inflections down as entries of their own. For instance, LDOCE lists *was* as a lexical entry, defined as "*1st and 3rd person sing. past tense of* BE". But although this solves the problem of retrievability, it is not the most elegant solution: it mixes lemmas and word-forms, and spreads related forms (are, is, was, being, are, am) around the dictionary. And it is has no clear demarcation criteria: should the Portuguese irregular 2nd person negative imperative *oiçais* of the verb *ouvir* (to listen) be included? Or the regular plural *sloegen* of the Dutch irregular past tense *sloeg* of the verb *slaan* (to hit)?

Access to the MorDebe database is primarily via the word-form: the user enters a word (string) he is looking for, and the database displays which form of which lemma it is. If the word appears in more than one inflectional paradigm, MorDebe lists all the lemmas the word belongs to. Since all the word-forms are explicitly listed, there is no difference in treatment between irregular forms, such as *was*, regular forms, such as *walked*, or cases where the inflectional form is identical to the citation form, such as the past tense *beat*.

A drawback of the inclusion of all inflected forms is that the set of word-forms becomes very large very quickly. MorDebe currently contains some 125.000 lemmas, but already slightly over 1,5 million word-forms for Portuguese, and it is growing steadily. This means that the possibility of browsing is virtually eliminated: MorDebe only provides access to the word-forms and lemmas via search queries. And despite the obvious advantages of searchable indexes, browseable lists have their own merits: people have a tendency of going through lists if they do not know exactly what they are looking for.

3.2. Storage vs. Computation

Storing all inflected forms explicitly is is not the most space-efficient way of storage: for the creation of the Portuguese data for MorDebe, a program was developed which generates verbal forms for all Portuguese verbs. In this applet, only the truly irregular forms are stored – all the rest, including transformations, are stored as rules. This applet implicitly contains all verb-forms in Portuguese, and stores them much more efficiently than the MorDebe database.

But although storage is less efficient, retrieval is much faster: to find all word-forms that are spelled as *walked*, a rule-based storage system would have to rely on morphological analysis to determine that it is the past tense of *walk*, whereas MorDebe can simply look up all the matching forms in the database. MorDebe even allows advanced search options, such as giving all word-forms ending on *-ked* or matching the pattern *wal**d*. In a rule-based system, these forms could only be retrieved by explicitly expanding all lemmas to find the matching word-forms.

Rule based system are effectively only useable in lemma-driven approaches, such as the CD version of the Houaiss dictionary (HoCD): when the user looks up a verb, the dictionary provides not only the normal definition, but also the full inflectional paradigm. But the access is always via the lemma. For a word-form driven system like MorDebe, explicit storage is the best solution.

4. Lexical Coverage

One of the main tasks of dictionary editors is lexical selection. Foremost, because the amount of lemmas presented in a dictionary is limited by physical boundaries, making a careful selection of the most relevant lemmas necessary. But also because it has to be assured that all lemmas included are well-established, correct, general language terms.

However, it is well known that the most frequent source of frustration of dictionary users is the absence of a word they are looking for. That is why Oppentocht & Schutz (2003) suggest that it might be useful to provide a much wider coverage in dictionaries, where possibly the words are not even supplied with definitions, since about 85% of all dictionary consultations are for checking spelling and word existence only.

This observation, although made from the perspective of dictionary design, describes much more the set-up of a lexical database than that of dictionary database. In the design of MorDebe, there are no reasons to reject words due to space limitations. And whereas the entries are intended to be adorned with semantic definitions, the lemma list and their definitions are stored separately, implying that it is not necessary for each lemma to have a semantic definition. When only the form and not the meaning is provided for a given lexical entry, it is nonetheless available for checking spelling and existence. Furthermore, the inflectional paradigm can still be provided on-line, making the system work as an orthographic guide. In that sense, MorDebe is exactly what Oppentocht & Schutz sketch as a future possibility.

As an orthographic guide, a lexical database only has real value if all recorded lemmas are thoroughly checked - if new words would be added too easily, the database reduces to an arbitrary list of words - likely to be correct, but not necessarily so. Therefore, new lemmas are only added to MorDebe after careful verification of existence and correctness. It is possible to add dubious words as well (and even incorrectly spelled words) but in MorDebe, this is only done when marking these lemmas explicitly as 'dubious' or 'wrong'. To ensure correctness of the database, MorDebe furthermore stores with each lemma its base of justification - either motivated by its occurrence in reliable dictionaries, or its occurrence in established sources - as well as when it was added and by whom.

Because of the virtual lack of limitations on number of lemmas, there is also a lessened restriction on generality of the term. Terminological words can be added to MorDebe, when explicitly marked as belonging to a specific terminological domain. In the interface, it is possible to restrict the search queries to only a specific domain, or only general language terms.

4.1. Neologisms

The inclusion of neologisms is a particularly difficult question in the design of

dictionaries, as for instance described by Agens (1995). On the one hand, users expect new words to be in their dictionaries, on the other hand, the inclusion of new words is a labour and cost intensive process, and there always is a respectable time-lag between the observation of neologisms by lexicographers, and their availability in the written end-product.

This is less so for lexical databases: the MorDebe database was set-up explicitly for the observation and description of neologisms using a web-based utility called NeoTrack (Janssen, *forthcoming*): daily, two major Portuguese newspapers are checked for possibly new words – i.e. words that are not in the MorDebe database. These neologism candidates are manually verified against corpora to verify whether they are real neologisms or already established words – and either added to a neologism database, or to the MorDebe database. Although it stays a labour intensive process, it is much easier to keep a lexical database up-to-date in this fashion than it is for the traditional dictionary database. And there is no delay between the observation of new words and their on-line availability in MorDebe.

5. Lemma Separation

A major issue on which there are differences between dictionary databases and MorDebe is the question of when to put two word-senses under the same lemma, and when to create different lemmas for them. In a dictionary database, lemma separation is always done in such a way to optimise both compactness and information, driven largely by semantic considerations.

The space limitation in dictionaries sometimes even leads to clustering of different lemmas under a single entry in the case of semantically transparent word-senses, as in the case of run-ons. Strictly speaking, run-ons are morphological derivation, listed at the end of a lemma, with only a grammatical category indication, and no semantic explanation, as the entry " **~ ly** *adv* " at the end of *royal* in the LDOCE dictionary - indicating that *royally* is the adverbial form of *royal* with the expected semantics, or the entry "**~ker**" at the end of *picnic* to indicate that a *picnicker* is someone who has a picnic.

In the case of zero-derivations, this clustering sometimes can go even further. The GDLP lists at the beginning of *beatão* (hypocrite) that it is either an adjective or a noun, as does PetRob for *réflexe* (reflex), clustering different word-classes under a single lexical entry. This clustering is a clear indication that lemma separation in dictionary databases is based primarily on semantic motivations.

5.1. Inflection based lemma separation

The central focus on inflections in MorDebe shifts the perspective on lemma separation - making it much more form-based. In MorDebe, the inflectional forms are seen as an integral part of the lemma. Therefore, two word senses with different inflections cannot be treated under the same lemma. So in MorDebe, there have to be two different lemmas for the verb *to ring,* because depending on its meanings, the past tense is either *rang* (phone) or *ringed* (bird). And there have to be two lexical entries for *band* in Dutch, since its plural can either be *banden* (tyres) or *bands* (bands).

However, within a single inflectional paradigm, alternative forms may occur: the past tense of the Dutch verb *waaien* (to blow) is either *woei* or *waaide*. And the plural of *pixel* (pixel) in Portuguese can either be *pixels* or *pixéis*. The question whether alternative inflectional forms lead to lemma separation is dependent on whether the two variant forms are intersubstitutable in all circumstances.

From an inflection-based perspective, it is clear that words of different word classes can never be listed under the same lemma: different word-classes have different inflectional paradigms. But taken very strictly, inflection based lemma separation goes even further: in its meaning of the celestial body circling the earth, *moon* is a *singulare tantum*. But in its poetic use as a synonym for *month,* or its more general meaning as a satellite body, it is not. And the word *foot* does not have a plural form in its use as a measurement unit. So strictly speaking, there should be two entries for *moon* and *foot*, one with a plural and one without. More extremely, the same would hold for all words that can be used both as mass nouns and count nouns.

This problem becomes even bigger if what Booij (1995) calls *inherent inflections* are taken into account: word-forms which are in a sense 'between' inflection and derivation. Traditionally, nominal gender is seen as inflectional in many Romance languages: the Portuguese word *geradora* is seen as an inflected form of *gerador* (originator). But female forms only exist for animate nouns. Consider the Portuguese word *amarelo* (yellow). In its base meaning as the colour, it does not have a plural, but in its more liberal sense of 'shade of yellow' it does. And when denoting someone with a yellow complexion ('pale person'), it even has a female singular and plural form. A strictly paradigm-based lemma system would require at least three different entries for *amarelo*: one without a plural, one with, and one with a female form as well.

To resolve these undesirable consequences of strict inflection-based lemma separation, MorDebe allows the existence of *semi-defectives*: words that have a defective inflection in some of their meanings. There is only one word *agua* (water) in MorDebe, which has a plural form *aguas*. And the fact that in its mass noun reading this plural cannot be used is seen as a semantic restriction imposed by a specific reading of the word.

It should be observed that these problems with inflection and word-senses can only be ignored in dictionaries because inflection is often not explicitly treated. But for instance the DLPC does explicitly list female forms, and for this reason, it is forced to view *amarelo* as homonymous, listing the colour and the pale person reading as separate entries.

6. Conclusion

Although there are many ways in which the design of lexical databases, or at least the MorDebe database, resembles the design of dictionary databases, this article shows that there are points in which they differ, due to their different purposes. Firstly, where dictionary databases have no real need for inflected word-form, they are crucial to the set-up of a lexical database. Secondly, where the focus in dictionary database is on selection to preserve compactness, consistency, and correctness, the emphasis in lexical databases is more on completeness and coverage. And thirdly, where lemma separation is almost exclusively governed by semantic issues in dictionary databases, it is largely driven by formal considerations in the MorDebe design.

All three of these differences lead to an increased amount of data in the LDB with respect to the dictionary database. This means that where the dictionary can still be browsed, MorDebe can only be used via search queries. On the other hand, where the dictionary can only be accessed via the lemma, MorDebe can be accessed via any word-form.

References

1. Dictionaries

OxRus: Paul Falla (ed.). 2000. *Oxford Russian-English Dictionary*. Oxford: OUP.

LDOCE: Randolph Quirks (ed.). 1987. *Longman Dictionary of Contemporary English, 2nd Edition*. Essex: Longman.

HoCD: Antônio Houaiss (ed.). 2001. *Dicionário Houaiss Eletrônico*. Lisboa, Rio de Janeiro: Círculo de Leitores.

CED: Patrick Hanks (ed.). 1986. *Collins Dictionary of the English Language, 2nd Edition*. London: Collins.

DLPC: João Malaca Casteleiro (ed.). 2001. *Dicionário da Língua Portuguesa Contemporânea da Academia das Ciências de Lisboa*. Lisboa: Verbos.

GDLP: Graciete Teixeira (ed.). 2004. *Grande Dicionário da Lingua Portuguesa*. Porto: Porto Editora.

PetRob: Paul Robert (ed.). 1989. *Le Petit Robert 1*. Paris: le Robert.

2. Other

Agnes, Michael. 1995. "Why It Isn't There: Practical Constraints on the Recording of Neologisms." *Dictionaries: Journal of the Dictionary Society of America*, vol. 16, p. 45 – 50.

Booij, Geert. 1995. Inherent versus contextual inflection and the split morphology hypothesis. *In:* Booij & van Marle (eds.) *Yearbook of Morphology 1995*. Dordrecht: Kluwer.

Janssen, Maarten. 2002. *SIMuLLDA: a Multilingual Lexical Database Application using a Structured Interlingua*. PhD Thesis, Utrecht University.

Janssen, Maarten. 2004. Multilingual Lexical Databases, Lexical Gaps, and SIM*u*LLDA. *International Journal of Lexicography*, vol. 17: 137 - 154.

Janssen, Maarten. *forthcoming*. Orthographic Neologisms. Selection criteria and semi-automatic detection. *Submitted to Terminology.*

Oppentocht, Lineke & Rik Schutz. 2003. Developments in Electronic Dictionary Design. *In:* P. van Sterkenburg (ed.) *A Practical Guide to Lexicography.* Amsterdam: John Benjamins Publishing.