Open Source Lexical Information Network

Maarten Janssen

ILTEC, Lisbon, Portugal

maarten@janssenweb.net

1. Introduction

Currently, there is a large number of lexical resources available: GENELEX, PAROLE, EuroWordNet and its follow-ups like GermaNet, MultiWordNet, etc. With this multitude of resources, the need arises for standardisation, in the guise of for instance the EAGLES, ISLE/MILE, EMELD, and TC37/SC4 projects. A current attempt in these standardisation framework is the conception of a network of lexical information, suggested as for instance the Semantic Web, LIRICS, the Open and Distributed Lexical Infrastructure (Peters, 2002), and the ELITE project (Calzolari, 2002). The overall aim of these frameworks is to improve interoperability, reusability, and availability through standardisation.

This article describes a proposal from a different perspective: the Open Source Lexical Information Network (henceforth OSLIN). The emphasis in OSLIN is on logistics rather than on structure; on attracting research groups to actually participate in the network and build actual resources, rather than on the structure of the participating databases. OSLIN puts forth a collection of principles which are aimed at assuring the actual creation of data in the network. Given its different approach, OSLIN should be seen as complementary to rather than an alternative for the standardisation projects mentioned above.

The OSLIN proposal is built around an infrastructure for semiautomatic neologism detection built at the ONP (Observatório de Neologia de Português) in Lisbon, consisting of a detection application (NeoTrack) integrated with a lexical database (MorDebe). Because of its design and purpose, the MorDebe database is large-scale, high-quality, up-to-date, easily accessible, and open-source. Furthermore, it is designed in a modular fashion, making it easy to link additional types of resources. This article will argue that these properties make MorDebe an ideal candidate for the creation of an open source lexical information network.

The article exists of three parts: the first part explains the general principles behind the OSLIN idea, optimising logistics in such a way to help attract actual resource data. The second part gives a brief overview of the MorDebe system, the set-up of its database, and the way it is integrated with NeoTrack application for the observation of neologisms. And the third part sketches how the MorDebe system could be extended to an open source lexical information network.

2. General Principles

A network of lexical resources is, by definition, dependent on the actual data it contains. Hence, an important aspect of a lexical network is the attraction of data, and assuring that these data not only exist, but are also available. However, there is a natural reluctance amongst research groups to share the fruits of their hard labour. The OSLIN proposal focusses on a number of principles to help overcome these hesitation, by making it interesting enough for research groups to share their data, and making it interesting to make them available as part of an open source lexical information network.

2.1 Shared Lexicon

The basic idea of the OSLIN proposal is that every lexical resource needs a lexicon - and preferably a lexicon of sufficient quality and size. In much linguistic research, the complexity of the compilation of a lexicon is often underestimated. The compilation of a high-quality lexicon requires precision, and hard labour, and well-defined standards, which have been the subject of a long tradition of lexicographic research.

An important concern of a open source lexical information network is hence to assure the co-operation of lexicographically capable research groups, willing to take care of the basic lexicon, and not only to set it up, but to keep it up-to-date, and make it fully open source.

2.2 Sharing Tools: Open-Source Infrastructure

Although it is relatively common to share information and making research results available, it is less common to share tools and utilities. But it is often the availability of a research infrastructure that makes it attractive for research teams to use a given framework. This is straightforwardly true for lexicographic research. Lexicology groups do not commonly have a strong programming team at hand. Hence, lexicographers are dependent on existing tools. The idea of the OSLIN network is to not only make the research data, but also the research tools and software.

The existence of a dedicated infrastructure for the semiautomatic detection of neologisms (i.e. NeoTrack and MorDebe) should make it interesting for lexicologists to use this existing framework rather than develop alternative software. Although some customisation will be required, the NeoTrack and MorDebe systems can be used for or at least extended to any language[1].

2.3 Sharing Resources

A large problem with lexical resources is that creating them is a labour intensive task, and as a result research groups are often reluctant to make the results of their hard work publicly avail-

[1] NeoTrack is based on standard UNIX tools, which are in principle ASCII based. With some small modifications, it will work for other alphabetic scripts as well, and the entire framework could relatively easily be changed to a Unicode set-up.

able, preferring rather to attempt to capitalise on the effort either in terms of money by selling it, or in terms of a research advantage by having the resources available for internal use only.

Hence one of the main tasks of a open source lexical information network is to assure that it is more advantageous for the participating teams to share their data than keep them private. The philosophy behind OSLIN is to first convince lexicographic teams to create large-scale, high-quality, maintainable resources, and convince other research by making this lexicon available for the construction of additional resources.

The motivation for lexicographers to use the OSLIN tools - and hence make their data open-source, is threefold: firstly, the availability of the dedicated NeoTrack system. Secondly, the fact that the main motivation of the research is not the lexicon itself, but the detection of neologisms. Since the lexicon itself is hence a spin-off rather than a primary goal, it should be easier to make it available, since the research teams involved can capitalise on the neologism data available in the project. And thirdly, the system comes with a web-consultation interface, allowing the general public to use the lexicon as an orthographic guide and inflection dictionary, assuring sufficient exposure.

2.4 Freedom of Design

One of the problems it that especially when taking semantic databases into account, there is no consensus on the proper type of lexical representation. From different angles, there is a wide array of lexical information which is relevant in one way or another for linguistic processing: hierarchical information, collocational data, frequency data, lexical fields, lexical proximity, sub-categorisation frames, dialectical information, usage data (style, specialisation, etc.), derivational relations, Qualia structures, etc. And most of these can be described in a multitude of theoretical frameworks.

Given the wide possible scope of lexical data, it is virtually impossible to impose standards on resources linked to the OSLIN network. This is not to say that research teams should be invited to disregard existing standards such as ISLE and MILE, but rather that lexical resources which are deviating from these standards could nonetheless be useful additions to the OSLIN network. The idea behind the OSLIN network is to have a very modular set-up, in which individual research groups can choose the internal organisation of their databases. In the long run, natural selection might select the most useful standards.

2.5 Division of Labour

Especially when talking about large scale lexical resources, no single research team is capable of creating a full set of resources. Furthermore, lexical resources span many different linguistic fields: lexicography, semantics, syntax, morphology, computer science, etc. Therefore, a full-scale lexical information network will require the co-operation of different research groups - each group with its own expertise. Given that these different groups will not share the same physical location, distributed research should be made possible.

In the OSLIN network, all data and tools are web-based (see section 4.2), allowing global co-operation on a single set of data. With the open source perspective on software and data, the software for the network can also be developed relatively independent from those responsible for the content. For instance, the construction of a query language (see section 4.3) would merit from a thorough investigation in terms of flexibility and efficiency.

3. ONP Set-up: MorDebe and NeoTrack

The ILTEC institute in Lisbon hosts a neologism observatory called the ONP, which aims at the detection and description of neologisms appearing in newspaper in European Portuguese. The ONP is part of NEOROM - a network of neologism observatories for all Romance language, co-ordinated by Teresa Cabré at the IULA in Barcelona. The ONP monitors two major Portuguese newspapers daily (the *Público* and the *Diário de Notícias*) for the occurrence of new words.

For the detection of neologism, a dedicated application was developed called *NeoTrack* (Janssen, *forthcoming*). NeoTrack is a web-based application based on the an exclusion-based neologism criterion: those words count as neologism candidates which do not appear in an exclusion list of known words. The system is semiautomatic, in the sense that the neologism candidate list has to be manually verified - manually judging all words on the list as neologisms, typos, name, existing words, etc. In the NeoTrack application, the MorDebe database is used both as the exclusion list for the automatic compilation of the list of neologism candidates, and for the storage of not previously encountered correct words.

3.2 Database Set-Up

The set-up of the MorDebe database is purposefully simple: it consists of little but two related tables, one listing all the lemmas, the other the list of all inflected forms for each lemma. The data are stored in a simple MySQL database format. The link between the two tables is implicit: every lemma is given a unique identifier, and every inflectional form relates to one of the records in the lemma table.

The complexity of the database is not in its global set-up, but in the treatment of the data - how to deal correctly with complexities of the lexicon such as:

1. homographs with different inflectional paradigms - such as the verb *to ring*, which can either inflect as *ring-rang-rung* (bell) or *ring-ringed-ringed* (bird)
2. variation within a single inflectional paradigm - such as the Dutch verb *waaien* (to blow), which has either *waaide* or *woei* as its past tense
3. dialect dependent inflections - such as the Portuguese verb *aguilhoar* (to spur), whose first person singular present tense is either *aguilhoo* (European Portuguese) or *aguilhôo* (Brazilian Portuguese)

4. orthographic variation (*medieval* vs. *mediaeval*) - which affects not only the citation form, but all inflected forms as well
5. meaning dependent inflection - such as the word Portuguese *verde* (green) which only has a plural in its meaning of *type of green*

Because the two tables in MorDebe are functionally independent, the consistency of the database has to be ensured externally: this is done by means of scripts to check whether there are inflections for all lemmas, whether every inflection refers to an existing lemma, whether all citation forms of the lemmas correspond to the significant inflectional form, etc. Furthermore, all maintenance is done using interface scripts, where the scripts make sure that the tables are treated consistently upon creation of new data, and modification or deletion of existing data.

Apart from the inflections, there are two other tables related to the lemma database: a grammatical database, and a derivational database. The grammatical database holds information on the grammatical behaviour of the lemmas: for verbs whether they can be used as transitive, intransitive, and/or reflexive verbs, and for nouns whether they are count and/or mass nouns. These data are stored separately since although they are relevant for lexicographic purposes, they are not taken to be identificational for lemmas, nor do they affect inflection[2].

The derivational database is used for the treatment of *inherent inflection* (Booij, 1995), which are in a sense between derivation and inflection. As an example: in MorDebe, male and female forms of Portuguese nouns are treated as separate lemmas. But traditionally, they are seen as inflectional variants. The derivational database links the lemmas *adjunto* (assistent) and *adjunta* (female assistent) as male/female alternation whereas still having them as separate lemmas.

Like the inflectional database, the grammar database and derivation database are linked implicitly to the lemma database by reference to the lemma ID. The administration environment of MorDebe also allows the management of these additional databases.

3.2 Size, Quality, and Up-to-Datedness

The fact that MorDebe is used as the exclusion list in neologism research, makes it necessarily a high-maintenance resource: it has to be large, it has to be high-quality, and it is by definition kept up-to date. The requirement of size comes from the fact that it is used for the computational analysis of actual language data, filtering the known words from the (potentially) new words. Hence, the list of known words should include as many correct words as possible, including those words that are commonly left out of dictionaries: inflected forms, semantically transparent terms, compound, low-frequency words and specialised terminology. The current size of the Portuguese MorDebe is about 125.000 lemmas, with around 1,5 million inflected forms.

---

[2] Reflexive verbs are in problem in this sense, since the reflexive pronouns are often taken to be part of the inflectional paradigm, especially in languages like Portuguese where these pronouns are expressed by means of clitics.

The requirements of quality comes from the lexicographic nature of the research: with too many arbitrary incorrect words on the exclusion list, the neologism criteria become blurred. But more importantly, without a strict verification of all added data, the database looses its value as an orthographic guide: the database is explicitly set-up as a multi-purpose database.

And the recentness of the database is a direct consequence of the use of the system: in the design of NeoTrack, newly encountered correct words are added directly to the MorDebe database, assuring that the database stays up-to date. Furthermore, the MorDebe database comes with on-line consultation page, where users can look up words and inflection, and in doing so provide comments and suggestion, helping to continuously improve the quality of the database.

4. OSLIN

The MorDebe design sketched in the previous section has a number of feature making it suitable as the core of a larger lexical network - extended not only in the sense of containing more different types of information for Portuguese, but also extend it to a multilingual network, with a parallel set-up for different languages.

In the extension of the MorDebe set-up to a larger network, a number of issues has to be addressed. This section discusses a number of problems and the way OSLIN is intended to deal with them. Since OSLIN currently is a proposal rather than an existing network, the solutions put forth in this section are open for debate.

4.1 Basic Design & Modularity

The idea behind the OSLIN network is to built a structured, distributed network of lexical information, organised around a central, lexicographically controlled lexicon. The coherence of the network originates from the reference to this common lexicon, in the way already indicated in section 3.2: every lemma in the lexicon has a unique record ID, and any external database is linked to the lexicon by reference this ID. Although this organisation is very low-key, it is only partially a loose structure - the resulting network is in another sense a single, relational database, where every related tables provides additional characterisations about the lemmas in the lexicon.

Because of the distributed nature of the organisation, the OSLIN network is very modular in set-up. At any point, any research group can create a new table with new data to the network, which can be maintained fully independently (barring cross-dependencies, see 4.4). There is even no objection against ‘competing’ databases within the network. As an example, one of the default tables of the network is a superficial grammatical table, specifying at least the traditional verb classes (transitive, intransitive, reflexive). But on top of this table, an alternative, much more elaborate grammatical characterisation could be added, providing for instance full sub-categorisation frames for verbs.

4.2 Web-Based Storage and Management

The OSLIN framework is completely web-based - all data are stored in plain MySQL databases, which can be accessed from anywhere over TCP/IP. Because of their availability, the web site on which the MorDebe database can be consulted does not provide access to an exported version of the database, but provides information directly from the database itself.

And even the entire management system is web-based: all editing of the MorDebe data is doing using browser-based forms, editing the data directly on the data-server. This method of direct remote editing has two advantages. Firstly, data are available immediately after they have been added, as well as any update or change. There is no publication delay, and the latest version of the database is always directly available.

And secondly, there are no local data involved, no local software, no dependency on local networks. This means that the MorDebe database can be edited by anybody from anywhere, as long as the computer is hooked up to the internet, and is equipped with a browser[3]. This means that there is no requirement for the researchers working on the database to share the same physical location.

4.3 Central Repository & Lexical Query Language

A distributed array of linked databases is only truly available if it is somehow centrally known which databases are available, and what their content is - without such information, the resources will be like unlinked web-pages: theoretically accessible, but only for those who know where to look. To that end, the idea is to set up a central server, providing meta-database information about the databases in the network.

Not only is this central server intended to provide general information, but also to provide an access point to the data themselves. The MorDebe database in its raw source does not comply with standards like MILE, mostly because the database has a simple tabular set-up rather than an XML-like organisation. Rather than forcing all data into an XML format, the idea in OSLIN is to have the central server which provides filters, translating the plain data gathered from the different databases into the desired format. To this end, a Lexical Query Language (LQL) is being developed which will provide an easily accessible way of gathering and translating the data - similar to for instance the EMELD Query Engine (Lewis *et al.*, 2001).

This method of on-the-fly translation has two additional advantages: firstly, the data can be translated into different standards if so desired. Despite the movement towards standardisation, there are still alternative standards in which the output data could be delivered. And secondly, although the standardisation is well on its way, there are still no real fixed standards. In case the standards were to be modified, only the LQL would have to be modified, whereas all the distributed databases can stay as they are.

---

[3] Of course, management is regulated by authorization, either simply by password, or even by IP blocking for sensitive data.

4.4 Semantic Entities and Polysemy

In its current guise, MorDebe does not have a semantic counterpart. But for a full-scale lexical information network, semantic entities are clearly necessary. The set-up of a semantic module in OSLIN is in principle straightforward: like the central list of lemmas, there is a list of word-senses, each with its own unique ID, to which other information can be linked. But there are a number of problem. Firstly, there is no citation form form word-senses. Although this problem can be solved by the introduction of glosses, as was done in the WordNet project, this is far from an ideal solution.

Secondly, concerning the central lexicon, there is not a lot of possible disagreement: depending on the application it might be better to include or exclude spelling deviations, and even disagreement about what the correct spelling of a word is, but these are marginal issues. For word-senses, this is far less straightforward. There are major differences in the number and nature of senses listed between different dictionaries, unclear differences between homonymy and polysemy, and as argued for instance by Pustejovsky (1995), even the very notion of sense enumeration is up for discussion.

4.5 Cross-Dependent Databases

Despite the modular set-up of OSLIN, there are clearly cases in which different tables are cross-dependent. To list some examples of cross-dependencies:

1. between meaning and inflection - the fact that the word *water* only has a plural in its sense of *body of water*.

2. between grammar and meaning - the fact that the verb *climb* is transitive in its meaning of *to go up*, but intransitive in its meaning of *to slope upward*.

3. between derivation and meaning - the fact that the female form of the French word *débiteur* is *débitrice* in its meaning of someone in debt, but *débiteuse* in its meaning of someone handling debits in a shop (Laporte, 1997)

The structurally most easy solution to these dependencies is to duplicate the entries in the relevant databases in all these cases: to have two entries for *water* because of its inflectional restrictions, to have a transitive and an intransitive entry for *climb*, and to have two entries for *débiteur* because of its derivation.

Although solving the problem at hand, this solution of reduplication has two drawbacks: firstly, it leads to an enormous increase in the number of lemmas. But secondly, modification in the lemma list to accommodate for dependencies with grammar or semantic tables would forego the modularity of the system - it would make the distributed maintenance of the individual databases very difficult.

For these reasons, OSLIN uses the principle of restrictive linking: in the semantic database, it will be possible to indicate that where *climb* as a lemma can be either transitive or intransi-

tive, in a given meaning it can be only one of the two - and that mass reading do not have a plural. Although this makes the dependencies between the different tables more complex, it keeps the tables separately maintainable.

4.6 Multi-word expressions

In the construction of a lexicon, multi-word expressions (MWE) form a substantial problem. On the one hand, there is a clear need for the inclusion of MWE's in dictionaries - many MWE's behave like single words in almost every sense of the word; cross-linguistically, there is no real difference between the English *book shop* and the Dutch *boekenwinkel*, except for the space between the words; and there is often even spelling variation within a single language: the Academia dictionary lists the Portuguese word *cônsul geral* (general consul) as two words, but the Porto Editora dictionary lists it as a compound: *cônsul-geral.*

But on the other hand, there is a gliding scale between fully fixed multi-word compounds such as *little finger* and free expressions such *little man*. And the proper characterisation of MWE's (and compounds) is more elaborate than that of single word lexemes: it is necessary to indicate how fixed the expression is, where in the compound it inflects, and what the word class of the constituent words is.

For this reason, in MorDebe we have opted to only include single word lexemes in the central lemma list, and to have a separate, more elaborate database for multi-word expression. This is merely a working hypothesis, since there is nothing in the set-up of the database that prevents the inclusion of MWE's. For Portuguese, this delimitation of the lemma list works rather well, but for other languages like English, it might be less suitable.

4.5 Cross-Language Dependencies

The construction of a multilingual network is clearly intended to allow the set-up of links between the various languages. There are various ways of setting up a cross-linguistic organisation (Janssen, 2004). Despite the advantages advocated in that article of the structured interlingua set-up of SIM*u*LLDA (Janssen, 2002), the modular design of OSLIN should allow for the coexistence of alternative linking mechanisms within the same lexical network.

5. Conclusion

In this article, I hope to have shown how and why the set-up of the MorDebe database could be extended to an open source lexical information network. To sum up the argument in favour of the OSLIN network:

1. The infrastructure of MorDebe and NeoTrack should attract the co-operation of lexicographic research teams for the construction and maintenance of large-scale, high-quality, open-source lexicons.

2. The existence of these large scale lexicons should attract the co-operation of other research teams to link additional linguistic resources

3. The modular set-up of the network should allow the different resources to be maintained separately, by different groups in distributed locations.

As is necessarily the case in the proposal for a lexical information network, the actual set-up will require the co-operation of a significant number of parties, require a substantial amount of time and money, and after its initial set-up, there is no guarantee that any additional interest will be attracted. But the relatively pragmatic, and logistically-oriented nature of the OSLIN proposal should at least assure its feasibility. And the existence of the Portuguese MorDebe database exemplifies the fact that the set-up of a large-scale, maintained database in this way is an attainable goal.

References

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

Calzolari, Nicoletta. 2002. *Language Resources & Semantic Web.* Presentation at *COLING-2002*, Taipei, 2002.

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 SIMuLLDA. *International Journal of Lexicography,* vol. 17: 189 - 194.

Janssen, Maarten. *forthcoming*. Orthographic Neologisms: selection criteria and semiautomatic detection. *Submitted to Terminology.*

Laporte, Éric. 1997. Les mots. un demi-siècle de traitements. *Traitement Automatique des Langues*, vol. 38: 47 - 68.

Lewis, William; Scott Farrar, and Terence Langendoen. 2001. Building a Knowledge Base of Morphosyntactic Terminology. *In:* Bird, Buneman and Liberman (eds.) *Proceedings of the IRCS Workshop on Linguistic Databases.* Philadelphia, 2001.

Peters, Wim. 2002. *Is this a way forward? Towards an Open and Distributed Lexical Infrastructure.* Presentation at *Constructing Lexica* workshop, Leiden, 2002.

Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge: MIT Press.