Analysis of Lexical Structures from Field Linguistics and Language Engineering

P. Wittenburg, W. Peters[+], S. Drude[++]

Max-Planck-Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
peter.wittenburg@mpi.nl
[+]University of Sheffield
[++]Free University of Berlin

Abstract

Lexica play an important role in every linguistic discipline. We are confronted with many types of lexica. Depending on the type of lexicon and the language we are currently faced with a large variety of structures from very simple tables to complex graphs, as was indicated by a recent overview of structures found in dictionaries from field linguistics and language engineering. It is important to assess these differences and aim at the integration of lexical resources in order to improve lexicon creation, exchange and reuse. This paper describes the first step towards the integration of existing structures and standards into a flexible abstract model.

1. Introduction

Lexica play an utterly important role in all linguistic sub disciplines ranging from Language Engineering to Field-Linguistics. The former generally deal with the main languages whereas the latter record minority and endangered languages. Lexica form an essential component in describing all relevant information about a language that can be associated with a structural unit of that language, e.g. a word, a morpheme, or even a whole sentence.

Lexica contain a wide range of linguistic information according to their nature and function. They vary from simple lists to complex resources with many types of linguistic information associated with the entries or elements. In general they can be of various types (the following list is not meant to be exhaustive): word list, machine readable dictionary, thesaurus, ontology, glossary, concordance, term bank, phonetic transcriptions, picture set, video shots, sound bits

Lexical resources are widely used for language and knowledge engineering. In both monolingual and multilingual environments, language resources play a crucial role in preparing, processing and managing the information and knowledge needed by computers as well as humans. In field-linguistics they also play a central role since they are focusing on basic linguistic units such as words, affixes and fixed expressions. The variety of lexical requirements in field linguistics is greater, since the language types differ widely.

Language technology components aiming at carrying out automatic parsing involve even more complex resources including dictionaries. In addition, multilingual dictionaries contain translation equivalents and concordances, and ontologies describe semantic relations between important concepts.

2. Formats and Structure Types

This large variety of available information and the linguistic differences between languages are the main reasons that there is a huge amount of different lexical structures and formats. Almost every lexicon comes along with its own specification that is defined by project and task requirements. The two terms "format" and "structure" cannot always be separated clearly. The term "structure" mostly refers to the internal organization of a document, while the term "format" addresses information which also has to do with the way information is presented to the user or stored by a computer program, which includes questions of data structure

Computer-based lexica come in various formats such as relational database format (which also implies the ER type of structure, see below), plain-text files in some proprietary format such as SHOEBOX[1] (which also has a typical structure, see below), MS WORD document formats and many others.

There are various ways in which textual and lexical data can be annotated and structured, depending on theoretical convictions and associated tools. The most widely used standards for the representation of structures are SGML, XML[2] and RDF [1]. But especially in field-linguistics we also meet special structure (and format) definitions such as from Shoebox, which basically has a feature-value pairs which can be embedded in tree structures. Since most of these field linguistic lexica are not meant to be processed automatically, but traditionally are meant to be put on paper, many of them are written in text processors such as MS WORD where the researchers are guided from the traditional structure (and format) principles of written lexica.

Data structures can take the form of typed feature structures such as Comlex[3] ([2]; see figure 1), relational tables, e.g. Celex[4] ([3] see figure 2), flat files (unnormalized relational format) or resource specific formats such as WordNet[5] [4] and EuroWordNet[6] [5]. The

---

[1] http://www.sil.org

[2] For introductions to SGML and XML see
http://msdn.microsoft.com/library/default.asp?url=/library/en-us/xmlsdk30/htm/xmtutxmltutorial.asp ,

http://www.projectcool.com/developer/xmlz/xmldtd/ ,
http://www.oasis-open.org/cover/xml.html

[3] http://cs.nyu.edu/cs/faculty/grishman/comlex.html

[4] see http://www.kun.nl/celex/

[5] see http://www.hum.uva.nl/~ewn

[6] see http://www.hum.uva.nl/~ewn

last two have been precompiled into binary and offset-based formats, i.e. optimized representations were chosen for operation. They come with tools for browsing and, in the case of WordNet, adding information and creating new WordNets.

```
(noun        :orth "assertion"  # orthography
             :subc ((noun-that-s) (noun-be-that-s)))
             # syntactic complementation
```

Figure 1: Comlex typed feature structure

The following example of the Celex Lexical Database[7] shows the morphological structure of the word '*abbreviation*'. The unique identifier expressed by the lemma number (lemmano) provides the key into orthographic, syntactic and phonetic information contained in different tables.

"*morphstatus: C*" means that the lemma is morphologically complex. "*imm1*" is one of the morphological analyses available in Celex, whereas "*formation*" expresses the rule on the basis of which this deverbal nominalization has been formed, in this case deletion of the final –*e* of the verbal root.

lemmano	lemma	morphstatus	Imm1	formation
26	abbreviation	C	abbreviate+ion	-e#

Figure 2: CELEX relational structure

The typical Shoebox structure very often used in field linguistics contains feature-value pairs embedded in tree structures in plain text files. An example is given in figure3.

```
\lx tan
 \lc tãtu
  \ps itr.v
    \ge run
        \pdl 1.sg inchoative
          \pdv atãnoko
  \ps tr.v
   \sn 1
    \ge paint
     \en to paint someboby or something with
colour
   \sn 2
    \ge write
     \xv atãnju op ete
        \xe I am writing on a paper
```

Figure 3: Shoebox type of feature value pairs

Increasingly often one can find lexica embedded in some relational database software, since the design interface is relatively simple and allows the user to easily create beautiful user interfaces. The structural basis is of course the same as for CELEX.

3. Lexical structures

To better understand the structural requirements of lexica it was decided to analyze a wide range of existing lexica and try to abstract from them to come to a more generic model. As was the case for the development of the Abstract Corpus Model which is the kernel of the

EUDICO tool set[8], the authors don't claim that there will be one "Generic Lexicon Model" which will fit all needs for all times, but we expect to be able to derive an Abstract Lexicon Model which has the expressional power to define a common framework for most of the lexica we know at this moment. A report was recently circulated with a few projects [6].

3.1. DOBES Lexica

With the help of a simple semi-graphical notation the lexical structures used in the DOBES project[9] were described. From the 8 documentation teams 11 different lexical structures could be identified. The most simple but very efficient for the intended documentation work were singular tables as spreadsheets or document files.

Tuvan orthography
Tuvan appendix
German orthography
Russian orthography
Russian appendix
Xakas orthography
Tofa orthography

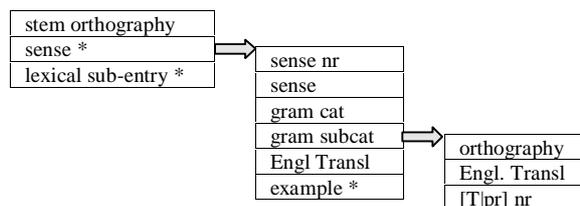Figure 4 shows the singular spreadsheet type lexicon used by the Tofa project within DOBES.

Figure 5 shows a part of one of the more complex lexica used in the Teop project within DOBES. A * sign stands for 1:n relations of sub-structures.
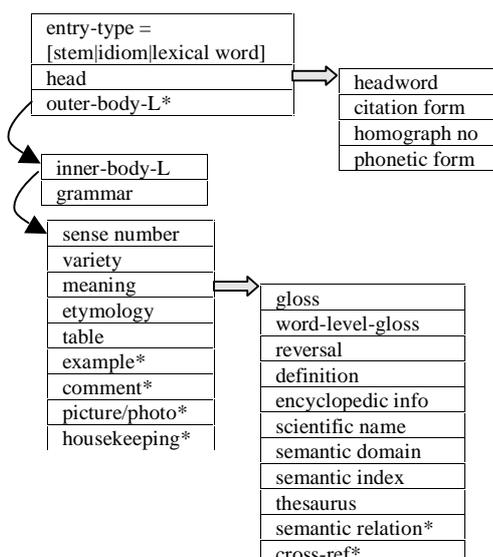
Figure 6 shows a small part of the complex structure worked out by the Aweti project within DOBES.

The most complex lexicon is set up by the Aweti team implemented as a complex hierarchy of Shoebox feature value pairs. The lexicon makes at high level a difference between 4 types of entries: entry-type = [stem | idiom | lexical word], entry-type = [auxiliary | inflectional affix], entry-type = [derivational word | derivational affix] or entry-type = [word form | allomorph]. For each type sub-structures exist. In the following example only an extraction of the first type is shown.

3.2. Lexica from Language Engineering

Beyond what was briefly indicated in chapter 2 the structural properties of a few other well-known lexica from language engineering were analyzed.

To be mentioned here is the GENELEX work the title of which claims to be generic. However, it was a concrete proposal for an exhaustive lexicon with definitions of structure and tag-sets. Its SGML structure consists of a huge DTD with specifications of three main layers (morphology, syntax, semantics) and many lexical elements integrated in tree-structures. GENELEX was used as a base line for the definition of the lexica from the PAROLE and SIMPLE[10] projects. These were an attempt to encode multilingual lexica in a uniform way with 12 fairly small sized example lexica as a result (see figure 7).

```
<MuS        id="V01015"
    %% morphological unit identifier%%
            gramcat="VERB"
            gramsubcat="MAIN"
    synulist="verb-cons-001V01015"    %%link to
            the syntactic units describing the
syntactic behavior of the entry%%
            autonomy="yes"
            combuf="UF1">
    <Gmu   naming="destroy"
                InP="Vinfl0">
    %%inflectional code%%
            <spelling>destroy</spelling>
    </Gmu>
</MuS>
```

Figure 7: PAROLE morphological entry

MULTILEX[11] was another project focusing on the implementation of 15 concrete lexica applying a structure derived from the EAGLES model of morphosyntactic annotation. Its data structure consists of three columns: wordform, lemma and morphosyntactic label. The latter provides a label for a number of classes. An example is:

adversities   adversity   Ncnp-
where adversities is a plural, neuter, countable noun.

The MILE (Multilingual Computational Lexicon) project recently started within ISLE has the task of standardizing multilingual lexica.

The early CELEX work was already described. It is realized as a rich set of relational tables for three languages where word form and lemma related information was separated.

3.3. Written Lexica

Also, examples from written dictionaries as analyzed by Bell&Bird [7] and Ide [8] were included to get a broad coverage. Bell&Bird studied more than 50 written lexica and found a number of characteristic organization principles and differences. The study showed mainly how the lexica differ with respect to
- the headword used and its characteristics
- the way senses are included

3.4. Other Lexica

Interesting proposals were made by two field researcher who focus on semantic relations between elements of lexical information. Schultze-Berndt [9] and colleagues implemented a lexicon by using the Hypercard mechanisms from Apple. She makes heavy use of semantic classes and also can create links from elements (words, set of words) in comment fields to other entries or elements within entries. In doing so she can realize complex semantic networks. Also Manning [10] stresses the relevance of supporting many different types of semantic relations between entries and attributes of entries. In his KirrKirr lexicon implementation he put much effort in visualizing these relations.

Although we did not find concrete lexica which make use of inheritance mechanisms, it is often reported that inheritance is a very important feature for computer-based lexica. So it is a structural requirement.

3.5. Summary

The analysis was in this stage not yet extended to lexica purely dedicated to cover semantic relations such as ontologies, thesauri etc., although some of the lexica discussed offer possibilities to use their structural possibilities to include such semantic relations.

As discussed above, the structure of the observed lexica varies considerably depending on the languages studied and the research interests. Simply structured dictionaries existing of a single table contrast with relational databases covering a large set of related tables. Also, many differences could be noticed with respect to the microstructure in dictionaries, i.e. the elements used to describe linguistic content and their underlying structural relations. This was supported by the observations found by Bell/Bird who showed, for example, that headwords and sense descriptions diverge.

The lexical structures found within the domains of language engineering and field linguistics diverge considerably. Between the two domains many similarities with respect to the requirements could be shown. Those attempts which use the term "generic" are not generic in the true sense. What GENELEX for example provides is an exhaustive list of tag sets which are embedded in a fixed hierarchical structure. This is not generic since the tag sets people are using differ largely, but especially since linguists differ largely with respect to the structural embedding of certain tags such as sense descriptions.

4. Standardization Efforts

[10] http://www.ub.es/gilcub/SIMPLE/simple.html
[11] http://www.ilc.pi.cnr.it/EAGLES96/lexarch

When discussing lexical structures it is important to review briefly the standardization work in the area of lexica and analyze in how they are relevant for structural issues. Much work has already been carried out on standardizing the description and creation of lexica, especially to facilitate language engineering applications.

While TEI[12] does not make detailed proposals for lexical tag sets, it does describe the structure of a dictionary entry in detail. Various standardization efforts such as EAGLES[13] and ISLE[14] worked out concrete proposals for standard lexical structures. GENELEX[15] can be seen as an early attempt to describe a generic lexicon structure with a complicated but exhaustive descriptive structure as was described above. As mentioned GENELEX was used to derive the lexica within the PAROLE and SIMPLE projects. Also MULTILEX was a standardization project, since it tried to work with a unified structure and tag set for several languages.

Partly within the area of terminology, other relevant standardization work was undertaken by the OLIF2 consortium (Open Lexicon Interchange Format)[16] resulting in the OLIF2 proposal. OLIF2 defines a large number of lexical features, but does not make statements about their structural embedding. Each OLIF2 entry is a monolingual entry containing various feature/value pairs, cross-references between entries in the same language lexicon, and transfers defining bilingual transfer relations. The OLIF2 proposal describes four main categories for features: administrative, morphological, syntactic, semantic. The features are similar to those found in other more generic lexicon proposals. Below are two examples with their descriptions:

PtOfSpeechDCS The ptOfSpeechDCS element (DCS is short for data category specification] holds data about a user-extended scheme for describing the part-of-speech of OLIF entries. Users can for example describe their additional part-of-speech tags by means of a URL or by means of CDATA sections.

SubjField The subjField element classifies the knowledge domain to which the lexical/terminological entry is assigned. Example values: agriculture, aviation.

MARTIF (Machine Reachable Terminology Interchange Format)[17] is another initiative in the area of terminology databases where especially a formal framework was worked out to define Data Categories - the basic elements of for example lexica. Such well-defined Data Categories will be available via open repositories.

Summarizing we can say that the standardizations were mainly on the level of definitions of data categories and tag sets. Some projects described structural layouts, but they are far away from being generic or even common enough to cover all lexical phenomena which were identified in the concrete lexica we analyzed.

5. Towards an Abstract Lexicon Model

Since almost every lexicon has its own idiosyncratic and inflexible format and structure it is difficult for the researchers and developers to easily access and combine them. On the other hand the analysis clearly indicates that it is possible to make abstractions from the concrete lexica and to define one underlying schema which all lexica we came across adhere to.

Recently, we found already comments which also go into this direction. Ide and Romary proposed a flexible formal model of dictionary structure and content on a workshop which was part of the MILE project in the ISLE initiative. This is also described in Ide et al [11]. The conceptualization of a dictionary as a tree is implemented by the CONCEDE lexical model [12]. Basically, a dictionary is seen as tree structure where the nodes can be associated with feature-value pairs. Inheritance mechanisms and cross-references allow them to build complex structures.

From the analysis and the papers found we can identify the structural phenomena which are necessary to formulate an Abstract Lexicon Model. We need

- simple building blocks which group a number of lexical attributes (data categories in the sense of terminology)
- a flexibility to associate labels and types with these attributes
- abstract data categories which refer to such building blocks (these references can be of type 1:N)
- inheritance mechanisms which indicate that attributes inherit characteristics from other attributes
- attributes which contain several elements (compounds, phrases, words) where each element can be addressed as a linguistic unit
- typed cross-references between attributes or elements of attributes

These simple mechanisms allow us to express all types of lexica which we came across until now. They cover the view of complex trees which lexical structures basically are. They also contain cross-references from descriptions or definitions within a lexical entry to descriptions of other entries, i.e. complex cross-reference structures where each cross-reference can have its own type. Finally they include inheritance mechanisms which describe operational characteristics of lexical attributes.

An implementation of an Abstract Lexicon Model can be based on frameworks such as UML (Unified Modeling Language) [13] or RDF (Resource Description Framework)[18]. The former has shown its expressional power in many software projects, while the latter offers a direct opening to the Semantic Web. Since RDF itself is not sufficient to express the mechanisms described above extensions will be necessary such as for example described in OntoMap [14].

6. References

[1] http://www.w3.org/RDF/

---

[12] http://www-tei.uic.edu/orgs/tei/

[13] http://www.ilc.pi.cnr.it/EAGLES96

[14] http://www.mpi.nl/ISLE

[15] http://www.ilc.pi.cnr.it/EAGLES96/lexarch

[16] http://www.olif.net/

[17] http://coral.lili.uni-bielefeld.de/~ttrippel/terminology/node76.html

[18] http://www.w3.org/RDF/

[2] Grishman, Ralph, Catherine Macleod and Adam Meyers (1994). *COMLEX Syntax: Building a Computational Lexicon*, Coling94, Kyoto

[3] Burnage (1990), Celex , a Guide for Users, Nijmegen, the Netherlands

[4] Fellbaum, Christiane (ed.) (1998), *WordNet: An Electronic Lexical Databas*e. Cambridge, Mass.: MIT Press.

[5] Vossen, P., Introduction to EuroWordNet. In: Nancy Ide, N., Greenstein, D. and Vossen, P. (eds), Special Issue on EuroWordNet. Computers and the Humanities, Volume 32, Nos. 2-3 1998. 73-89.

[6] Wittenburg, P. (2001) Lexical Structures. MPI Technical Report. MPI Nijmegen

[7] J. Bell, S. Bird (2000) A Preliminary Study of the Structure of Lexicon Entries. Paper presented at the Workshop on Web-Based Language Documentation and Description. Philadelphia.

[8] Ide, N., Le Maitre, J., and Veronis, J.(1991), *Outline for a Model of Lexical Databases.* RIAO91, Barcelona

[9] Schultze-Berndt, E. (2001) Unpublished Manuscript of a contribution to a lexicon workshop. MPI Nijmegen

[10] KirrKirr Lexicon: www.sultry.arts.usyd.edu.au/kirrkirr

[11] Ide, N., Kilgarriff, A. and Romary, L. (2000), *A Formal Model of Dictionary Structure and Content*, Euralex, Stuttgart

[12] Erjavec, T., Evans, R., Ide, N., Kigarriff, A. (2000), The Concede Model for Lexical Databases, LREC, Granada

[13] Booch, G., Rumbaugh, J. and Jacobson, I. (1999), The Unified Modelling Language User Guide. Addison Wesley Longman

[14] A. Kiryakov, K. Simov, M. Dimitrov. OntoMap: The Upper-Ontology Portal. In: Proceedings of "Formal Ontology in Information Systems", FOIS-2001, October 17-19, 2001, Ogunquit, Maine**.**