1. Introduction

Language resources 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 this document we concentrate on a particular kind of language resources, the lexical resources. In general they can be of various types (the list below is not 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. Relevant research areas include computational linguistics, computerized lexicography, and language engineering. Computational linguistics and language engineering provide the methodology for preparing, recording, processing and reusing the language resources. Computerized lexicography supplies the tools for the efficient preparation and processing of lexical data. Language engineering provides us with the tools for representing, managing and accessing knowledge mediated by linguistic data with different degree of complexity. Language resource management cannot be efficient without a strong language engineering component.
2. The representation format of linguistic resources

There are various ways in which textual and lexical data can be annotated and structured, depending on theoretical convictions and associated tools. An enumeration of the main types of data structure encountered is given in section 2.3. The most widely used standards for resource representation are SGML, XML and RDF, shortly described in the following section.

2.1 SGML and XML

These are widely used standards for annotating text structure. XML has superseded SGML, but there is a wide-spread availability of resources in SGML format. For several SGML and XML tutorials/information pages see footnote1. In 1994 the Text Encoding Initiative (TEI)2 published a set of detailed recommendations for the encoding and transcription of many types of written and spoken materials, using an extensible SGML framework. This format has also been influential in lexicon creation projects such as PAROLE and SIMPLE as well as in defining EAGLES and related standards (see section 3).

The following TEI example comes from its guidelines for encoding print dictionaries3 shows a dictionary entry that provides information on several aspects of orthography, phonology, syntax and semantics.

```xml
<entry>
  <form>
    <orth>competitior</orth>  orthography
    <hyph>com|peti|tor</hyph>  syllabification
    <pron>k@m`petit@(r)</pron>  pronunciation
  </form>
  <gramGrp>
    <pos>n</pos>    part of speech
  </gramGrp>
  <def>person who competes.</def>  definition
</entry>
```

This TEI specification for dictionaries has been adopted and extended within the CONCEDE project (Consortium for Central European Dictionary Encoding)4.

2.2 RDF

The Resource Description Framework (RDF)5 is, as its name implies, a framework for describing and interchanging metadata. It provides a model and a syntax s for metadata so that independent parties can exchange it and use it.

---

http://www.projectcool.com/developer/xmlz/xmlldtd/
http://www.oasis-open.org/cover/xml.html
http://www.oasis-open.org/cover/general.html
http://www.w3.org/MarkUp/SGML/
2 http://www.tei-c.org/
3 http://www.tei-c.org/Guidelines/DI.htm
4 http://www.itri.bton.ac.uk/projects/concede/
5 http://www.w3.org/RDF/
At the core, RDF data consists of nodes and attached attribute/value pairs. Nodes can be any web resources (pages, servers, basically anything for which you can give a Universal Resource Identifier (URI)), even other instances of metadata. Attributes are named properties of the nodes, and their values are either atomic (text strings, numbers, etc.) or other resources or metadata instances. In short, this mechanism allows us to build labeled directed graphs which can be converted into XML. For a tutorial see footnote

An example is shown below, where the attribute creator attached to the resource uniquely identified by the URI has the value John Smith.

```xml
<RDF:RDF>
  <RDF:Description RDF:HREF = "http://URI-of-Document">
    <DC:Creator>John Smith</DC:Creator>
  </RDF:Description>
</RDF:RDF>
```

Different linguistic classification systems will provide different packages of resource/properties/values combinations. These packages are called vocabularies. RDF in itself does not contain any predefined vocabularies for authoring metadata (see section 3).

### 2.3 Main types of data structure

1) **Typed feature structures:**

A feature structure is composed of pairs of attributes (called features) and their values, which can also be seen as partial functions from features to values.

Each lexical entry is organized as a list of categorized features. Each list consists of a type symbol followed by zero or more keyword-value pairs. Each value may in turn be an atom, a string, a list of strings, feature-value list, or a list of feature-value lists. For a more detailed introduction we refer the reader to Shieber (1986)\(^7\).

An example is the Comlex Syntax database\(^8\):

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

2) **Relational format:**

A relational database consists of a set of relations between entities. Each role in that relation is called an attribute. Conceptually, a relation is a table whose columns correspond to attributes, and each row, or tuple, specifies all the values of attributes of a given entry. Attributes have only atomic value, that is, values which cannot be decomposed. In other words, each row-to-column intersection contains one, and only one, value.

The following example of the Celex Lexical Database\(^9\) 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.

---


\(^8\) [http://cs.nyu.edu/cs/faculty/grishman/comlex.html](http://cs.nyu.edu/cs/faculty/grishman/comlex.html)
“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.

<table>
<thead>
<tr>
<th>lemma</th>
<th>morphstatus</th>
<th>formation</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>abbreviation</td>
<td>abbreviate+ion -e#</td>
</tr>
</tbody>
</table>

The following example taken from LDOCE shows one possible conversion of a printed dictionary entry (abandon) into relational format.

abandon 1 /ˈbændən/ n [T1] 1 to leave completely and for ever; desert: The sailors abandoned the burning ship. 2 to leave (a relation or friend) in a thoughtless or cruel way: He abandoned his wife and went away with all their money. 3 to give up, esp. without finishing: The search was abandoned when night came, even though the child had not been found. 4 (to) t o give (oneself) up completely to a feeling, desire, etc.: He abandoned himself to grief | abandoned behaviour. -- ~ment n [U].

abandon 2 n [U] the state when one’s feelings and actions are uncontrolled; freedom from control: The people were so excited that they jumped and shouted with abandon/in gay abandon.

The derived relational database has four tables. Each table expresses dependency of the value(s) of one or more columns on a set of key columns.

The names of the columns are the following:

- HW = headword
- PS = part of speech
- HN = homograph number
- SN = sense number
- DF = definition text
- EX = example
- GC = grammar code
- PR = pronunciation

DEFINITION (Key: HW, PS, HN, SN)

<table>
<thead>
<tr>
<th>HW</th>
<th>PS</th>
<th>HN</th>
<th>SN</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>1</td>
<td>to leave completely and for ever</td>
</tr>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>1</td>
<td>desert</td>
</tr>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>2</td>
<td>to leave (a relation or friend) in a thoughtless or cruel way</td>
</tr>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>3</td>
<td>to give up, esp. without finishing</td>
</tr>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>4</td>
<td>to give (oneself) up completely to a feeling, desire, etc.</td>
</tr>
<tr>
<td>abandon</td>
<td>N</td>
<td>2</td>
<td>0</td>
<td>The people were so excited that they jumped and shouted with abandon/in gay abandon</td>
</tr>
</tbody>
</table>

‘0’ is used for entries that have only one sense and no explicit numbering in the paper based entry (see above).

9 http://www.kun.nl/celex/
PRONUNCIATION (Key: HW, PS, HN)

<table>
<thead>
<tr>
<th>HW</th>
<th>PS</th>
<th>HN</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>/bænd n/</td>
</tr>
<tr>
<td>abandon</td>
<td>N</td>
<td>2</td>
<td>/bænd n/</td>
</tr>
</tbody>
</table>

EXAMPLE (Key: HW, PS, HN, SN)

<table>
<thead>
<tr>
<th>HW</th>
<th>PS</th>
<th>HN</th>
<th>SN</th>
<th>EX</th>
</tr>
</thead>
<tbody>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>1</td>
<td>The sailors abandoned the burning ship</td>
</tr>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>2</td>
<td>He abandoned his wife and went away with all their money</td>
</tr>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>3</td>
<td>The search was abandoned when night came, even though the child had not been found</td>
</tr>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>4</td>
<td>He abandoned himself to grief</td>
</tr>
<tr>
<td>abandon</td>
<td>N</td>
<td>2</td>
<td>0</td>
<td>the state when one's feelings and actions are uncontrolled</td>
</tr>
<tr>
<td>abandon</td>
<td>N</td>
<td>2</td>
<td>0</td>
<td>freedom from control</td>
</tr>
</tbody>
</table>

CODE (Key: HW, PS, HN, SN)

<table>
<thead>
<tr>
<th>HW</th>
<th>PS</th>
<th>HN</th>
<th>SN</th>
<th>GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>1</td>
<td>T1</td>
</tr>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>2</td>
<td>T1</td>
</tr>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>3</td>
<td>T1</td>
</tr>
<tr>
<td>abandon</td>
<td>V</td>
<td>1</td>
<td>4</td>
<td>T1</td>
</tr>
<tr>
<td>abandon</td>
<td>N</td>
<td>2</td>
<td>0</td>
<td>U</td>
</tr>
</tbody>
</table>

In the last table the value of grammar code is dependent on the first four columns. This means that the grammar code may change from one word sense to another, which is quite often the case in the dictionary.

3) Resource specific format. This class accounts for resource or company specific data structures that mostly come with access routines or interfaces. Examples of these are WordNet\(^{10}\) that uses data files indexed on byte offsets, and EuroWordNet\(^{11}\) that has its own specific import and export format: of which the following is an example:

The actual names of the features or columns (e.g. “orth” in Comlex) and the nature of the associated values (e.g. “-e#” in Celex) constitute the resource specific vocabulary of the linguistic metadescription. On top of that, different resources describe the same type of linguistic information by means of different terms (e.g. “orth” vs “lemma”) or divide up the conceptual space into different chunks of different granularity (compare the Ldoce and WordNet syntactic subcategorization information; see below).

\(^{10}\) [http://www.cogsci.princeton.edu/~wn/](http://www.cogsci.princeton.edu/~wn/)

\(^{11}\) [http://www.hum.uva.nl/~ewn/](http://www.hum.uva.nl/~ewn/).
3. Standardization

If one wants to describe language resources, create new ones, or make efficient use of them one needs the appropriate methodology, methodology standards, software tools, and the respective standards for mark-up, interchange, exploitation and evaluation.

Much work has already been carried out on standardizing the description and creation of lexicons, especially to facilitate language engineering applications. While TEI\(^\text{12}\) 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\(^\text{13}\) and ISLE\(^\text{14}\) worked out concrete proposals for standard lexical structures. GENELEX\(^\text{15}\) can be seen as an early attempt to describe a generic lexicon structure with a complicated but exhaustive descriptive structure. The PAROLE and SIMPLE\(^\text{16}\) projects were an attempt to encode multilingual lexicons in a uniform way with 12 fairly small sized example lexicons as a result. MULTILEX\(^\text{17}\) was another project focusing on the implementation of 15 concrete lexicons applying a structure derived from the GENELEX model. The MILE (Multilingual Computational Lexicon) project [8] recently started within ISLE has the task of standardizing multilingual lexicons.

Partly within the area of terminology, other relevant work was undertaken by the OLIF2 consortium (Open Lexicon Interchange Format)\(^\text{18}\) 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 for features describes four main categories: administrative, morphological, syntactic, semantic. The features are similar to those found in other more generic lexicon proposals. Below are a few examples with their descriptions:

\textbf{PtOfSpeechDCS} \quad 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.

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

\textbf{SubjFieldDCS} \quad The subjFieldDCS element holds data about a user-extended scheme for describing the subject field information of OLIF entries (see the comment for the ptOfSpeechDCS element for more information).

\textbf{Syllabification} \quad The syllabification element holds data about the syllable boundaries within the entry string. Example use: do-cu-men-ta-ry, li-be-ra-li-ty.

---

12 \url{http://www-tei.uic.edu/orgs/tei/}
13 \url{http://www.ilc.pi.cnr.it/EAGLES96}
14 \url{http://www.mpi.nl/ISLE}
15 \url{http://www.ilc.pi.cnr.it/EAGLES96/lexarch}
16 \url{http://www.ub.es/gilcub/SIMPLE/simple.html}
17 \url{http://www.ilc.pi.cnr.it/EAGLES96/lexarch}
18 \url{http://www.olif.net/}
**SyllabificationMarkInfo** The syllabificationMarkInfo element holds data about editorial practice adopted with respect to syllabification in the original. Example use: we use ‘*’ as marker.

**SynFrame** The synFrame element classifies the syntactic frame for the entry string (subcategorisation). Example values: subj-imps-opt, dobj-opt.

**SynFrameDCS** The synFrameDCS element holds data about a user-extended scheme for describing the syntactic frames of OLIF entries (see the comment for the ptOfSpeechDCS element for more information).

**SynPosition** The synPosition element classifies the unmarked positioning of the entry string syntactically. Example values: prenoun, cl-init.

**SynStruct** The synStruct element holds data about the constituent structure of a multiword entry string (note the possibilities provided for single words by means of the morphStruct element). Example use: [[adj][noun]] (General Ledger).

Much work has been done in the area of terminology databases. The MARTIF (Machine Reachable Terminology Interchange Format)\(^\text{19}\) work describes a format to facilitate the interchange of terminological data among terminology management systems. This work resulted in the ISO 12200 specifications\(^\text{20}\). Complementary to that ISO 12620 specifies how “Data Categories” (the basic elements for describing lexical content) have to be defined. Term related information specifies the linguistic type of the terms. This is done by assigning linguistic attributes to the entries such as part of speech (cf. OLIF\(^2\) above). Descriptive information links the terms to domains and points to positions in concept hierarchies. Administrative and proprietary information can also be added to each term such as creator name and creation date.

The SALT project (Standards-based Access to Lexicon and Terminologies)\(^\text{21}\) was recently initiated mainly driven by the needs from language engineering. SALT suggests the XLT (XML representations of Lexicons and Terminologies)\(^\text{22}\) family of formats for representing, manipulating, and sharing terminological data. The core structure of SALT is based on the MARTIF proposal.

### 4 Metadata for lexical description

The information linguistic categories contain and their structural relations can, for reasons of generality and conceptual clarity, best be described by means of metadata, i.e. information about the types of available linguistic information. In fact, all information contained in lexicons and related resources is metadata but the resources differ in terms of terminology, level of granularity of linguistic description and data format (see section 2.3). This is exemplified by the standardization efforts described above (e.g. the term ptOfSpeechDCS is OLIF specific and may contain different values from e.g. EAGLES) and the comparison of resources below in section 5. Metadata are being proposed by initiatives such as OLIF and ISLE. Their function is to describe and access resources in a standard fashion. The ISLE consortium\(^\text{23}\) has issued a draft proposal\(^\text{24}\)

---

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


21 [http://www.ttt.org/salt/](http://www.ttt.org/salt/)

22 [http://www.ttt.org/oscar/xlt/DXLT.html](http://www.ttt.org/oscar/xlt/DXLT.html)

that divides lexical metadata up into two main groups: external (information about the lexicon as a whole) and internal (information about the lexical entry).

4.1 External metadata

External units of information describe the lexicon as an object and can be the following (see the draft proposal for a full list):

- **Name** A short name which identifies the Lexicon
- **Title** A more elaborated title of the Lexicon
- **Date** Date of the creation and major modifications
- **Version** Version indication
- **Creator** The responsible persons who created the resource
  - **Name** Name of creators
  - **Contact** block of features related to contact person or organization (see below)
  - **Description** A suitable description associated with the set of creators
- **Project** A block to describe the project
  - **Name** Short name of the project
  - **ID** Unique project identifier
  - **Contact** contact address sub schema
  - **Description** some space for descriptions to be associated with the project
- **LexiconType** Type following some taxonomy (see e.g. list in Section 1)
- **Object Languages** A block to describe the languages included in the lexicon
  - **Description** some space for a prose description
  - **MultilingualityType** languages can occur in different flavors in lexica, they can occur as multilingual entries in ML lexica, but they also can occur as translations of for example sense descriptions; this difference can be indicated with the help of a controlled vocabulary
  - **Language** a list of languages included, each language be described in a substr ucture
  - **Format** a rough indication of the format the lexicon is in such as relational table, structured plain text, some XML format, html format, ...
- **AccessTool** many lexicons are only interpretable via concrete access tools such as Shoebox, ORACLE, FoxPro, Access, Web-Browser,…
- **Media** this entry tells whether the lexicon includes audio or video samples or graphics
- **Character Encoding**+ this list should give an impression of the type of fonts needed to render all data included such as UTF-8, ISO-latin
- **Size** the size of the lexicon in bytes
- **No Lexical Entries** the number of lexical entries the lexicon includes
- **Access sub-schema where access info is given (see below)**
- **Keys** a possibility to add feature/value pairs to define new keywords
- **Source** this entry describes which sources were used to build the lexicon
- **References** block to cover references to publications etc.

Access
- **ResourceLink** URL pointing to the resource if it is directly accessible
- **Availability** codification of terms of access (has to be worked out
- **Description** prose description associated with access
- **Date** date of statements about access
- **Owner** defines the owner of the lexicon
- **Publisher** defines the publisher of the lexicon
- **Contact** specifies a sub-schema describing whom to contact

Contact
- **Name** name of the contact person
- **Address** address info
- **Email** email address
- **Organization** name of an institution
- **Language**
  - **Language ID** formal language specifier from ISO or SIL lists
  - **Name** general name of the language
  - **Description** a description of the language can be associated

---

4.2 Internal metadata

This type of data gives us information about the linguistic content of the lexicons. The following linguistic units of description have been distinguished within the ISLE lexical metadata initiative. This list is not meant to be exhaustive.

**Modality** indicates which mode of communication is captured in the lexicon. Possible values are:
- Spoken
- Written
- Sign

**Headword type** indication of the linguistic nature of the entry in the lexicon. Possible values are:
- Sentence
- Phrase
- Wordform
- Lemma entry conforming to the unmarked wordform (e.g. infinitive for verbs).
- Abstract Lemma entry not conforming to any wordform of the group subsumed by the lemma.
- Stem
- Affix

**Orthography** possible values are:
- Hyphenated Spelling
- Syllabified Spelling
- Spelling Variants - orthographic variations with or without preferred spelling information
- Citations

**Morphology** possible values are:
- Stem deep or surface stem
- Stem Allomorphy variations at stem level
- Segmentation analysis into morphological constituents such as affixes
- Production rules governing the production of surface forms on the basis of stems
- Typology any classification of entries or morphological entities

**Morphosyntax** possible values are:
- Part of Speech syntactic class of the entry.
- Inflection any inflectional or conjugational information
- Countability pluralization properties
- Gradability e.g. adjectival comparative/superlative constructions
- Gender e.g. neuter
- Typology any classification of entries

**Syntax** possible values are:
- Complementation Syntactic complementation
- Alternation alternative complementation patterns
- Modification e.g. adjectival modification patterns
- Shallow Parsing segmentation into chunks
- Deep Parsing finer grained analysis below chunk level
- Functional Parsing syntactic functions such as subject
- Collocations significant juxtaposed entries/wordforms
- Typology any classification, e.g. prepositional/phrasal verb

**Phonology** possible values are:
- Transcription any type of phonetic/phonological transcription
- IPA Transcription transcription in International Phonetic Alphabet
- CV pattern transcription in terms of consonant-vocal combinations
- Constituent Structure segmentation into phonetic constituents
- Intonation stress marking, constituent length etc.

**Semantics** possible values are:
- Sense distinction polysemy and/or homonymy
Another parallel initiative within ISLE, the EAGLES/ISLE Working Group for Multilingual Lexicons, aims at the standardization of multilingual lexical entries. For this purpose a checklist has been created that has much overlap with the metadata set listed above, but is in many cases more fine-grained in its coverage. A brief explanation of abbreviations used: SL= source language; TL = target language; IPA=International Phonetic Alphabet

<table>
<thead>
<tr>
<th>Entry component</th>
<th>Information content</th>
<th>Mode</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 headword</td>
<td>lexical form(s) of the headword: how the headword is spelt</td>
<td>SL</td>
<td>Helps both SL and TL users find the information they are looking for</td>
</tr>
<tr>
<td>2 Phonetic transcription</td>
<td>how the headword (or variant form etc.) is pronounced (in <em>International Phonetic Alphabet</em>)</td>
<td>IPA</td>
<td>Helps user pronounce the word correctly</td>
</tr>
<tr>
<td>3 variant form</td>
<td>alternative spelling of headword or slight variation in the form of this word</td>
<td>SL</td>
<td>helps both types of user find the information they are looking for</td>
</tr>
<tr>
<td>4 inflected form</td>
<td>other grammatical forms of the lemma (headword)</td>
<td>SL</td>
<td>helps dec user find the information they are looking for</td>
</tr>
<tr>
<td>5 Cross-reference</td>
<td>indication of another headword whose entry holds relevant information, or some other part of the dictionary where this may be found</td>
<td>code</td>
<td>helps both types of user find the information they are looking for, or other useful information</td>
</tr>
<tr>
<td>6 Morphosyntactic information</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://www2.arts.gla.ac.uk/IPA/ipa.html
<table>
<thead>
<tr>
<th></th>
<th>Part-of-speech marker</th>
<th>part of speech of the headword (or the secondary headword)</th>
<th>code</th>
<th>helps both types of user find the information they are looking for, by focussing the search</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>Inflectional class</td>
<td>Inflectional paradigm of the entry</td>
<td>code</td>
<td>helps SL user use TL item correctly helps TL user disambiguate TL word helps TL user use SL item correctly helps SL user disambiguate SL word</td>
</tr>
<tr>
<td>c</td>
<td>Derivation</td>
<td>Cross-part-of-speech-information, morphologically derived forms</td>
<td>SL</td>
<td>helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</td>
</tr>
<tr>
<td>d</td>
<td>Gender</td>
<td>Information about the gender of the entry in SL and TL</td>
<td>code</td>
<td>helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</td>
</tr>
<tr>
<td>e</td>
<td>Number</td>
<td>Information about the grammatical number of the entry in SL and TL</td>
<td>code</td>
<td>helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</td>
</tr>
<tr>
<td>f</td>
<td>Mass vs. Count</td>
<td>Information whether the a noun is mass or count, in SL and TL</td>
<td>code</td>
<td>helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</td>
</tr>
<tr>
<td>g</td>
<td>Gradation</td>
<td>For adverbs and adjectives</td>
<td>code</td>
<td>helps SL user use TL item correctly helps TL user disambiguate TL word</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th></th>
<th>Subdivision counter</th>
<th>indicates the start of new section or subsection ('sense')</th>
<th>number / letter</th>
<th>'signpost' helping user to find their way about the entry more efficiently</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Entry subdivision</td>
<td>separate section or subsection in entry (often called <em>dictionary sense</em>)</td>
<td>Dictionary text</td>
<td>breaks up entry, making it easier to read and find what is being sought</td>
</tr>
<tr>
<td>9</td>
<td>Sense indicator</td>
<td>synonym or paraphrase of headword in this sense, or other brief sense clue indicating specific sense of SL or TL item</td>
<td>SL</td>
<td>helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</td>
</tr>
<tr>
<td>10</td>
<td>linguistic label</td>
<td>the style, register, regional variety, etc. of the SL or TL item</td>
<td>code</td>
<td>helps SL user identify the sense of the headword helps both users translate helps TL user understand</td>
</tr>
<tr>
<td>11</td>
<td>Syntactic information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>Subcategorization frame</td>
<td>(i.) Number and types of complements (ii.) syntactic introducer of a complement (e.g. preposition, case, etc.) (iii.) type of syntactic representation (e.g. constituents, functional, etc.) etc.</td>
<td>code</td>
<td>helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</td>
</tr>
<tr>
<td>b</td>
<td>Obligatory of complements</td>
<td>Information whether a certain complement is obligatory or not</td>
<td>code</td>
<td>helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</td>
</tr>
<tr>
<td>c</td>
<td>Auxiliary</td>
<td>Which type of auxiliary is selected by a given predicate (in certain languages auxiliary selection is related to issues like unaccusativity, which on turn lie at the interface between lexicon and syntax)</td>
<td>code</td>
<td>acts as a sense indicator helps SL user select appropriate TL equivalent</td>
</tr>
<tr>
<td></td>
<td>Light or support construction</td>
<td>Constructions with light verbs</td>
<td>SL or TL</td>
<td>helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>e</td>
<td>Periphrastic constructions</td>
<td>Constructions containing periphrasis, usage, semantic value, etc.</td>
<td>SL or TL</td>
<td>helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</td>
</tr>
<tr>
<td>f</td>
<td>Phrasal verbs</td>
<td>Particular representation of phrasal constructions</td>
<td>SL or TL</td>
<td>helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</td>
</tr>
<tr>
<td>g</td>
<td>Collocator</td>
<td>(i.) typical subject/object of verb, noun modified by adjective etc. (ii.) type of collocation relation represented etc.</td>
<td>SL or TL</td>
<td>acts as a sense indicator helps SL user select appropriate TL equivalent helps TL user translate or understand the SL item</td>
</tr>
<tr>
<td>h</td>
<td>Alternations</td>
<td>Syntactic alternations an entry can enter into</td>
<td>Code</td>
<td>acts as a sense indicator</td>
</tr>
</tbody>
</table>

12 Semantic information

<table>
<thead>
<tr>
<th></th>
<th>Semantic type</th>
<th>Reference to an ontology of types which are used to classify word senses</th>
<th>Code</th>
<th>helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>Argument structure</td>
<td>Argument frames, plus semantic information identifying the type of the arguments, selectional constraints, etc.</td>
<td>Code</td>
<td>helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</td>
</tr>
<tr>
<td>c</td>
<td>Semantic relations</td>
<td>Different types of relations (e.g. synonymy, antonymy, meronymy, hyperonymy, Qualia Roles, etc.) between word senses, etc.</td>
<td>Code</td>
<td>acts as SL sense indicator for SL user acts as TL sense indicator for TL user</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Regular polysemy</td>
<td>Representation of regular polysemous alternations</td>
<td>Code helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Domain</td>
<td>Information concerning the terminological domain to which a given sense belongs</td>
<td>Helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Decomposition</td>
<td>Representation of relevant meaning component, e.g. causativity, agentivity, motion, etc.</td>
<td>Code acts as SL sense indicator for SL user acts as TL sense indicator for TL user</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Translation</td>
<td>TL equivalent of SL item</td>
<td>TL helps TL user understand helps both users translate</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Gloss</td>
<td>TL explanation of meaning of an SL item which has no direct equivalent in the TL</td>
<td>TL helps TL user understand helps both users translate</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>near-equivalent</td>
<td>TL item corresponding to an SL item which has no direct equivalent in the TL</td>
<td>TL helps TL user understand helps both users translate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>example phrase (straightforward)</td>
<td>a phrase or sentence illustrating the non-idiomatic use of the headword, in a context where the TL equivalent is virtually a word-to-word translation</td>
<td>SL acts as SL sense indicator for SL user acts as TL sense indicator for TL user helps TL &amp; SL users to use the foreign-language item correctly</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>example phrase (problematic)</td>
<td>a phrase or sentence illustrating a non-idiomatic use of headword in a context where a specific TL equivalent is required (i.e. an SL example which is easily understandable for the TL speaker, but presents translation problems for the SL speaker)</td>
<td>SL helps SL user avoid a translating error acts as a sense indicator for SL user helps TL user subsequently to use the SL item correctly</td>
<td></td>
</tr>
</tbody>
</table>
18. multiword unit (idiomatic) multiword expression (MWE) containing the headword (*the term MWE covers idioms, fixed & semi-fixed collocations, compounds etc.*)

19. Subheadword also secondary headword

20. usage note

21. Frequency

<table>
<thead>
<tr>
<th>18</th>
<th>multiword unit</th>
<th>(idiomatic) multiword expression (MWE) containing the headword (<em>the term MWE covers idioms, fixed &amp; semi-fixed collocations, compounds etc.</em>)</th>
<th>SL</th>
<th>helps both users translate</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>Subheadword also secondary headword</td>
<td>lemma morphologically related to the headword, figuring as head of a sub-entry (<em>subheadwords can be compounds, phrasal verbs, etc.</em>)</td>
<td>SL</td>
<td>saves space helps both types of user find the information they are looking for</td>
</tr>
<tr>
<td>20</td>
<td>usage note</td>
<td>how the headword is used; ‘macro’ information which cannot appear at every appropriate entry; warning of cultural differences between the two languages; etc.</td>
<td>SL or TL</td>
<td>helps both types of user to avoid misunderstandings about the foreign language item, based on own-language knowledge</td>
</tr>
<tr>
<td>21</td>
<td>Frequency</td>
<td>Information about the frequency of the entry</td>
<td>code</td>
<td>helps both users translate</td>
</tr>
</tbody>
</table>

5. Comparison of resources using metadata

In order to get an impression of the usefulness of metadata for comparison and evaluation of lexical resources the table below lists the linguistic content of four resources according to a metadata set that is a general level subset of the ISLE sets discussed above. The resources under examination are the Longman Dictionary of Contemporary English (LDOCE)\(^{26}\), The Celex database, WordNet and the Cambridge International Dictionary of English (CIDE)\(^{27}\).

<table>
<thead>
<tr>
<th>ORTHOGRAPHY</th>
<th>LDOCE</th>
<th>CELEX</th>
<th>WORDNET</th>
<th>CIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spelling</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Spelling variants</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Syllabification</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Capitalisation</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PHONOLOGY</th>
<th>LDOCE</th>
<th>CELEX</th>
<th>WORDNET</th>
<th>CIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonetic transcription</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Stress marking</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>


**MORPHO-SYNTAX**

| Part of speech | 1 | 1 | 1 | 1 |
| Inflection | 1 | 1 | 1 | 1 |
| Conjugation | 1 | 1 | 1 | 1 |
| Countability | 1 | 1 | 0 | 1 |
| Gradation (e.g. busy, busier) | 1 | 1 | 1 | 1 |
| Type (e.g. common noun, auxiliary verb) | 1 | 1 | 0 | 1 |
| Gender | 1 | 1 | 0 | 1 |

**MORPHOLOGY**

| Derivation/composition | 0 | 1 | 0 | 0 |
| Segmentation | 0 | 1 | 0 | 0 |

**SYNTAX**

| Alternation | 1 | 1 | 1 | 1 |
| Complementation | 1 | 1 | 1 | 1 |
| Positional (attributive, predicative) | 1 | 1 | 0 | 0 |
| Analysis of multi-word-units | 0 | 0 | 0 | 1 |
| Collocational restrictions | 0 | 0 | 0 | 1 |

**SEMANTICS**

| Senses | 1 | 0 | 1 | 1 |
| Ontological classification | 1 | 0 | 1 | 1 |
| Semantic relation | 1 | 0 | 1 | 1 |
| Definition | 1 | 0 | 1 | 1 |
| Preference | 1 | 0 | 1 | 1 |
| Regular polysemy | 0 | 0 | 1 | 0 |
| Domain | 1 | 0 | 0 | 1 |
| Idiom | 1 | 0 | 0 | 1 |

**OTHER**

| Usage notes | 1 | 0 | 0 | 1 |
| Examples | 1 | 0 | 1 | 1 |
| Translation | 0 | 0 | 0 | 0 |
| Frequency | 0 | 1 | 0 | 0 |

In order to refine the comparison the high level information provided by this classification system it can be extended by choosing increasingly fine-grained levels of linguistic description by e.g. incorporating the complete ISLE checklists and more. For example, a subclassification of multi-word-units can be provided on the basis of their constituent parts (fixed phrases, compounds, idioms, support verb constructions, phrasal verbs). Verb complementation can be further subdivided into (in-/di-)transitive, copula, phrasal verb, prepositional verb and support verb. Maximum refinement is obtained when the linguistic information has been decomposed into the most basic information units. The result is a very complex structure of highly interrelated blocks of minimal linguistic information, and is exemplified by the GENELEX architecture.

An example of differences between resources and how they fit into the metadata classification scheme is the encoding of verbal complementation and preference information in the four resources mentioned above. For this particular type of linguistic content the following pieces of information are found:
CELEX:
- complementation: set of labels expressed by boolean values in columns:
  e.g. transitive: Yes; ditransitive: No
- no semantic preference

LDOCE:
- complementation: codes representing verb classes.
  e.g. D1: ditransitive
- preference: 34 semantic classes for subject/object/indirect object slots
  e.g. liquid/movable

WordNet:
- complementation and preference: surface patterns
  ‘Somebody Vs somebody something’

CIDE:
- complementation: codes representing verb classes.
  e.g. ‘T’: transitive
- preference: 40 semantic classes for subject/object/indirect object slots
  e.g. ‘human’/‘clothing’

The figure below illustrates the case of the verb “fall”.

---

Syntax      Semantics

Part of speech
(WN,Ldoce,CIDE) Celex: ‘class’

Verb

verbal complementation
WN: ‘verb pattern’
Ldoce: ‘grammar code’

Intransitive

LDOCE: ‘I*’
CELEX: ‘Intransitive: yes’
WordNet: frame ‘Somebody/Something Vs’
CIDE: ‘I’

subject preference
Ldoce: subj. pref.
CIDE: ‘selpos’

WordNet:
‘Somebody/Something’
LDOCE: ‘Z’ (unmarked)
CIDE: ‘Human/Object’
‘FALL’

Complementation and preference for the verb ‘fall’

Linking resource specific units of description to a fine-grained metamodel will have several advantages:
- it enables direct access to the information contained in the resource in a uniform way
- it provides an explicit link between representation formats. This enables the integration of different resource-specific attributes/values:
  ‘I*’ = ‘I’ = ‘intransitive’ = ‘somebody Vs’
  word form <derived from> lemma <synonym> headword
  <morphologically_composed_of> stem