

Analysis of Lexical Structures from Field Linguistics and Language Engineering

Wim Peters - University of Sheffield

Sebastian Drude - University of Berlin

Peter Wittenburg - Max-Planck-Institute

Content

1. Introduction
2. Format and Structure Types
3. Lexical Structures
 1. DOBES Lexicons
 2. Language Engineering Lexicons
 3. Printed Lexicons
 4. Other
4. Standardization Efforts
5. Perspectives

Introduction

- lexicons are central in language processing since they store all information associated with structural units in languages
- appear in great variety depending on a.o. nature and function (word lists, machine readable dictionaries, thesauri, ontologies, glossaries, concordances, term bank, ...)
- lexical resources are relevant
 - for HLT to be able to automatically process texts
 - for Field Linguistics to discover the repeating structural units, their embedding and meaning
- mostly lexicons are multilingual - be it that they include sense descriptions in another than the target language

Large Variety

- huge amount of different lexical structures and formats
 - differences between languages
 - differences in purpose and content
 - differences in linguistic theory
- each project comes with its own specifications
no way to combine, to use the same tools, to easily read
- structure = related to internal organization of the lexicon
- format = includes aspects of presentation and representation
- often not separable

Principal question

Is there a minimal set of structural units of linguistic description to cover most lexicons?

Formats and Structure Types

Structures:

- relational tables (CELEX, many small lexicons)
- spreadsheets
- typed feature structures (Comlex)
- feature value pairs embedded in trees (Shoebox)
- traditional printed dictionaries
- SGML tree structure (Genelex)
- resource specific & optimized (WordNet)

Formats:

- | | |
|--------------------------------------|-------------------------------------|
| • relational DB | CELEX, many small lexicons |
| • idiosyncratic labeling of AV pairs | Shoebox |
| • MS Word | traditional way - close to printout |
| • MS Excel | simple mechanisms |
| • SGML/XML | |
| • ... | |

Structures in DOBES

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

simple spreadsheet

stem orthography
sense *
lexical sub-entry *

little more complex incl 1:N relations

sense nr
sense
gram cat
gram subcat
Engl Transl
example *

orthography
Engl. Transl
[T pr] nr

entry-type = [stem idiom lexical word]
head
outer-body-L*

headword
citation form
homograph no
phonetic form

small part of a complex lexicon structure at top level 4 different entry types (only one is shown)

inner-body-L
grammar

sense number
variety
meaning
etymology
table
example*
comment*
picture/photo*
housekeeping*

gloss
word-level-gloss
reversal
definition
encyclopedic info
scientific name
semantic domain
semantic index
thesaurus
semantic relation*
cross-ref*

Structures in LE Lexicons 1

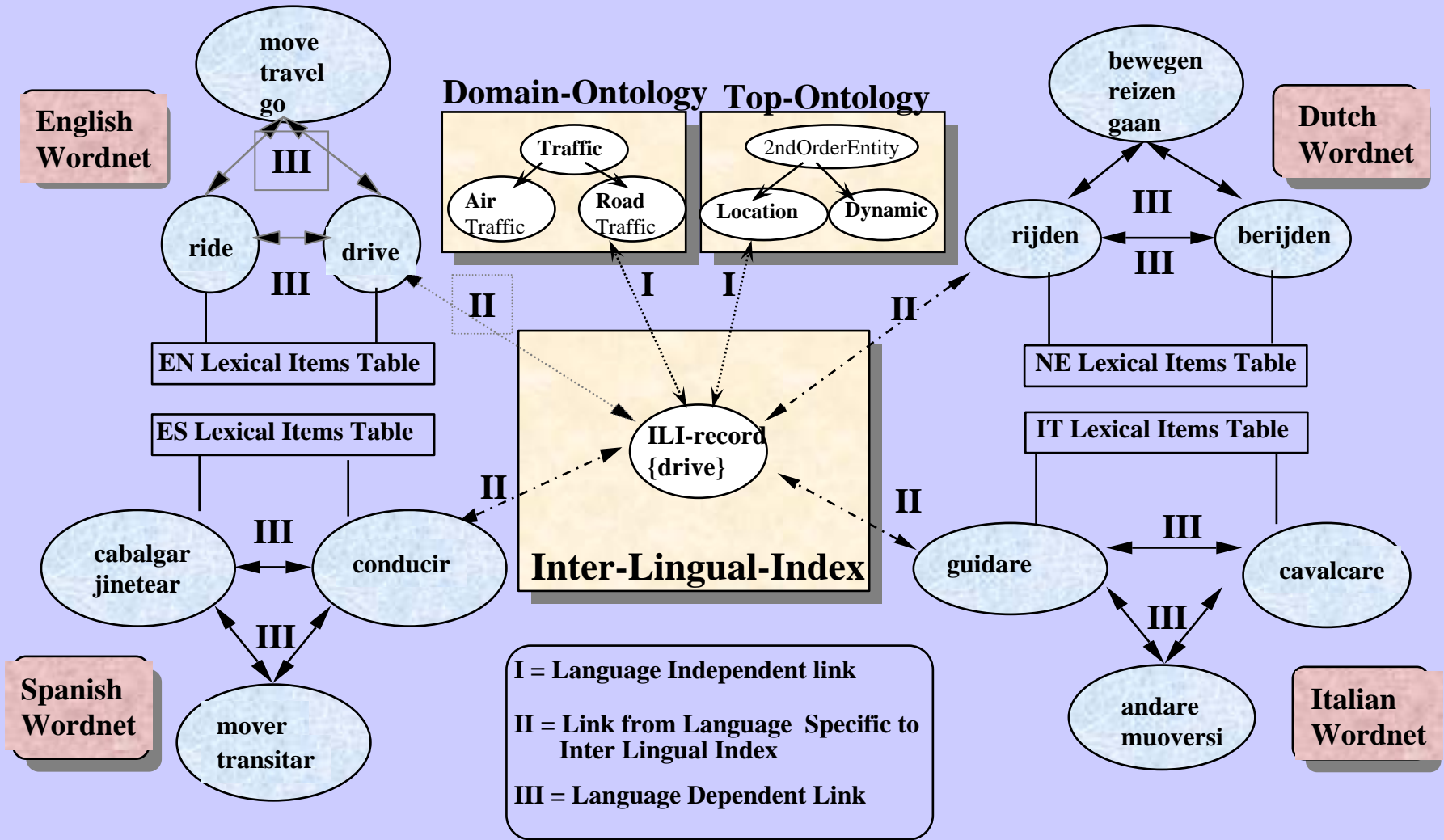
- GENELEX
 - 31 pages of SGML DTD with tag specifications
 - 3 major dimensions (morphology, syntax, semantics)
 - all attributes part of fixed tree
- PAROLE and SIMPLE based on GENELEX
 - attempt to encode multilingual lexicons (12) in a uniform way

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

Structures in LE Lexica 2

- MULTILEX
 - 15 lexicons applying EAGLES model of morphosyntax
 - simple three column structure (wordform, lemma, ms label)
- CELEX - large lexicons for E/D/G
 - relational approach with lemma and wordform tables
 - lemma: orthography, phonology, morphology, syntax
 - wordform: orthography, phonology, morphology
 - implicitly tree structure
- EuroWordNet
 - Multilingual thesaurus/MRD containing 8 languages
 - Language-specific independent wordnets
 - Wordnets represent unique concept lexicalization patterns in 8 languages, based on sense-inventories of mono- and bilingual dictionaries
 - Words are grouped according to synonymy (synsets)
 - Semantic relations between synsets (e.g. hyponymy, meronymy, antonymy)
 - Conceptual equivalence relations with interlingua

Architecture of the EuroWordNet Data Base



Printed Lexicons

- Bell & Bird (examination of about 50 written lexicons)
 - microstructure differs (language, linguistic theory)
 - different ways to mark linguistic information
 - no recursions for sub entries (assume finite no. of recursion)
 - no clear distinction between head and body information
 - headword is in general either orthographic or phonetic form
 - sense definitions sometimes are references to other entries
- problem with printed lexicons:
 - presentation format is equal to representation format
 - in fact underlying representation can be more simple

Other Lexicons

- Schultze-Berndt (and others) developed a Hypercard lexicon that enables the creation of semantic networks by cross referencing words in comment fields
- Manning created the KirrKirr lexicon
 - representation and visualization of semantic relations

Standardization Efforts

- TEI: description of dictionary entry in detail
 - EAGLES: morphosyntactic classification
 - ISLE: metadata descriptions of lexicons
 - Genelex: exhaustive tag set in fixed tree structure
 - PAROLE/SIMPLE: subschemes from Genelex
 - Multilex: uniform lexicons based on EAGLES
-
- OLIF2: definition of lexical features, no structure
 - MARTIF: formal framework for data category definition
-
- much work on the definition of data categories and tag sets
 - some definitions of structural layouts - but limited

Abstract Lexicon Model

Is there a minimal set of structural units of linguistic description to cover most lexicons?

- analysis until now suggests common underlying conceptual schema in terms of the uniformity of units of linguistic description
- In line with Ide & Romary on dictionaries:
 - The structure of each resource reflects a tree with nodes
 - nodes are associated with feature-value pairs
 - need inheritance mechanisms and cross references
- ALM building elements
 - simple building blocks that group lexical attributes
 - associating labels and types with these attributes
 - abstract data categories which refer to such building blocks
 - inheritance mechanisms to inherit characteristics from attributes
 - attributes containing several elements where each is a linguistic unit
 - typed cross-references between attributes or elements of attributes

Flexibility of description

- Minimum amount of dependencies between data categories
- Resource specific dependencies between data categories are outside model
- Local or central addition of new data categories if needed
- Data categories superset of all descriptive units used in resource creation/management

Need for:

- broad discussion to determine data categories and dependencies
- A mechanism for the evaluation of new user/resource-based data categories for their inclusion in ALM

What would we gain?

- New lexicons can be built with the help of a standardised set of descriptive units
- Towards pooling of terminology across areas of linguistic R&D activity
- Interoperability and interchangeability made possible
- ALM is superset of linguistic knowledge and therefore covers multimodality and multilinguality
- Addability of new data categories ensures user autonomy
- Resources can be accessed in a flexible way
- Resources can be compared and evaluated on the basis of the ALM and a specification of the resource-specific dependencies between the ALM elements