Lexicon Formalisms
An Example of a Lexical Standard

Kurt Eberle

April 21 2016
Outline

Objectives of LMF

LMF Philosophy

The LMF ‘Core Package’

LMF Extensions

Benefit of LMF
Lexical Markup Framework (LMF)

- Slides/Content by Ulrich Heid (Hildesheim), Thorsten Trippel (Tübingen)
- Objectives of LMF – why standardize lexical modeling?
- LMF philosophy:
  core + extensions
- Basic features:
  Coding - data categories - modeling language
- LMF-“Core Package”:
  Content – Applicability – Examples
- Extensions:
  Principles – Examples
- Benefit and usability of LFM
- Summary – Outlook
Outline

Objectives of LMF

LMF Philosophy

The LMF ‘Core Package’

LMF Extensions

Benefit of LMF
Lexical representation – Objectives of LMF

⇒ Lexicon/Lexicography: one of the relevant core areas
Lexical representation – Objectives of LMF

- “Business environment” of ISO TC-37/SC4: computational linguistics – computerized lexicography – language engineering ⇒ Lexicon/Lexicography: one of the relevant core areas
- language resources for “language and knowledge engineering”: “..., lexica, ontologies and terminologies...”: multilingual, multimedia, multimodal
Lexical representation – Objectives of LMF

  ⇒ Lexicon/Lexicography: one of the relevant core areas

- Language resources for “language and knowledge engineering”: “…, lexica, ontologies and terminologies…”: multilingual, multimedia, multimodal

- Objectives of LMF (Sect. 1: Scope):
Lexical representation – Objectives of LMF

⇒ Lexicon/Lexicography: one of the relevant core areas
▶ language resources for “language and knowledge engineering”: “..., lexica, ontologies and terminologies...”: multilingual, multimedia, multimodal
▶ Objectives of LMF (Sect. 1: Scope):
  – Meta model (suggestion of modeling methods)
Lexical representation – Objectives of LMF

⇒ Lexicon/Lexicography: one of the relevant core areas
▶ language resources for “language and knowledge engineering”: “..., lexica, ontologies and terminologies...”: multilingual, multimedia, multimodal
▶ Objectives of LMF (Sect. 1: Scope):
  – Meta model (suggestion of modeling methods)
  – for the representation of data in lexical data collections
Lexical representation – Objectives of LMF

  ⇒ Lexicon/Lexicography: one of the relevant core areas
- Language resources for “language and knowledge engineering”: “…, lexica, ontologies and terminologies…”: multilingual, multimedia, multimodal
- Objectives of LMF (Sect. 1: Scope):
  - Meta model (suggestion of modeling methods)
  - For the representation of data in lexical data collections
  - Used in monolingual or multilingual computational applications
Goals of LMF

Delimitations with respect to ISO-1951

- LMF relates to lexical information to be used in applications:
  Dictionaries as data structures
Goals of LMF

Delimitations with respect to ISO-1951

- LMF relates to lexical information to be used in applications:
  Dictionaries as data structures
    - Support for combination of lexica, for integration of existing lexica, for development of new types of lexica
Goals of LMF

Delimitations with respect to ISO-1951

- LMF relates to lexical information
to be used in applications:
  Dictionaries as data structures
  - Support for combination of lexica,
    for integration of existing lexica
    for development of new types of lexica
  - Mechanisms to enable representation of existing dictionaries,
    “as far as possible”, for instance GENELEX, EAGLES,
    ISLE/MILE
Goals of LMF

Delimitations with respect to ISO-1951

- LMF relates to lexical information to be used in applications:
  Dictionaries as data structures
  - Support for combination of lexica, for integration of existing lexica for development of new types of lexica
  - Mechanisms to enable representation of existing dictionaries, “as far as possible”, for instance GENELEX, EAGLES, ISLE/MILE

- ISO-1951 relates to dictionaries, which are printed (at least as an option) and used by humans:
  dictionaries as texts with macro structure and micro structure
Infrastructural embedding of LMF

Coding – data categories – modeling tools

- LMF is unicode based
Infrastructural embedding of LMF
Coding – data categories – modeling tools

- LMF is unicode based
- LMF supports all languages coding with ISO language codes
Infrastructural embedding of LMF

Coding – data categories – modeling tools

- LMF is unicode based
- LMF supports all languages coding with ISO language codes
- LMF is a meta model
Infrastructural embedding of LMF

Coding – data categories – modeling tools

- LMF is unicode based
- LMF supports all languages coding with ISO language codes
- LMF is a meta model
- LMF is specified using UML: UML diagrams
UML used in the specification of LMF

Some details of the UML language (1/2)

- Packages
  - Partial models
  - may be dependent on other packages, packages can be combined

![Diagram showing package dependence]

- Classes
  - with names
  - with links to other classes
  - with texts describing the use
  - but: without definition of the attributes: only examples!
UML used in the specification of LMF

Some details of the UML language (1/2)

- **Packages**
  - Partial models
  - may be dependent on other packages, packages can be combined

- **Classes**
  - with names
  - with links to other classes
  - with texts describing the use
  - but: without definition of the attributes: only examples!
UML used in the specification of LMF

Some details of the UML language (2/2)

- classes vs instances

```
; class A
```
UML used in the specification of LMF

Some details of the UML language (2/2)

- classes vs instances
- standard relations between classes
  - association
  - aggregation (e.g. part / whole )
  - generalization (mother vs daughter)
Outline

Objectives of LMF

LMF Philosophy

The LMF ‘Core Package’

LMF Extensions

Benefit of LMF
LMF philosophy: core and extensions
Core package und 8 extensions (1/2)

- Core Package
  - NLP morphology extension
    - NLP paradigms of inflection
    - NLP multiword patterns
  - NLP syntax
  - NLP semantics
  - NLP-multilingual descriptions
  - Machine-Readable-Dictionary extension
  - Constraint extension
LMF philosophy: core and extensions
Core package and 8 extensions (2/2)
Outline

Objectives of LMF

LMF Philosophy

The LMF ‘Core Package’

LMF Extensions

Benefit of LMF
The LMF ‘Core Package’

Graphical Representation
The LMF ‘Core Package’

Entries in the dictionary

- ≥ 1 Entry in 1 lexicon
- ≥ 1 Lexicon in resource
- Each resource has 1 description:
  - language coding
  - optionally administrative information
The LMF ‘Core Package’

Entries

- An entry has $\geq 1$ forms
- An entry has $\geq 0$ senses

- The form can have $\geq 0$ variants
- Senses may be structured internally
- Senses may have definitions represented by texts
The LMF ‘Core Package’
How plausible and flexible is the model?

> **Entries are always stored in one dictionary:**
> What about representations where several dictionaries are merged?
> Example: *Duden GWDS und Duden Universal*
The LMF ‘Core Package’

How plausible and flexible is the model?

- *Entries are always stored in one dictionary:* What about representations where several dictionaries are merged? Example: *Duden GWDS und Duden Universal*

- *Each entry has at least one form:* What about
  - Terminology data basis with entries for concepts?
  - GermaNet’s hypothetical concepts, for instance “hierarchical teacher”?
  - Synsets as semantic abstraction?
The LMF ‘Core Package’

How plausible and flexible is the model?

- **Entries are always stored in one dictionary:**
  What about representations where several dictionaries are merged?
  Example: *Duden GWDS* und *Duden Universal*

- **Each entry has at least one form:**
  What about
  - Terminology data basis with entries for concepts?
  - GermaNet’s hypothetical concepts, for instance “hierarchical teacher”?
  - Synsets as semantic abstraction?

- **Entries may omit senses/be without semantic content:**
  - lexica for morphology, pronunciation, syntax
  - Rechtschreibduden (orthographical information)
Outline

Objectives of LMF

LMF Philosophy

The LMF ‘Core Package’

LMF Extensions

Benefit of LMF
Extensions of the LMF core

Example Syntax (1/4)

In general:

▶ all extensions are optional

▶ extensions are added/related to the class
  “Lexical Entry” - or to the description of “Sense”
Extensions of the LMF core

Example Syntax (1/4)

In general:

- all extensions are optional
- extensions are added/related to the class “Lexical Entry” - or to the description of “Sense”

Syntax package

- “Syntactic Behaviour”: highly general attribute
Extensions of the LMF core

Example Syntax (1/4)

In general:

▶ all extensions are optional
▶ extensions are added/related to the class
  “Lexical Entry” - or to the description of “Sense”

Syntax package

▶ “Syntactic Behaviour”: highly general attribute
▶ “Subcategorization Frame” provides possibility of annotating
  “syntactic Argument”
▶ “Subcategorization Frame” allows mapping to semantics

⇒ Flexible, very general representation,
  no decision about syntax theory
Extensions of the LMF core
Example Syntax (2/4)
Extensions of the LMF core

Example Syntax Syntax (3/4)
Extensions of the LMF core

Example Syntax (4/4)

```xml
<LexicalEntry>
    <feat att="partOfSpeech" val="verb"/>
    <Lemma>
        <feat att="writtenForm" val="amare"/>
    </Lemma>
    <SyntacticBehaviour subcategorizationFrames="regularSVOAvere"/>
</LexicalEntry>

<SubcategorizationFrame id="regularSVOAvere">
    <LexemeProperty>
        <feat att="auxiliary" val="avere"/>
        <feat att="position" val="1"/>
    </LexemeProperty>
    <SyntacticArgument>
        <feat att="syntacticFunction" val="subject"/>
        <feat att="syntacticConstituent" val="NP"/>
    </SyntacticArgument>
    <SyntacticArgument>
        <feat att="syntacticFunction" val="object"/>
        <feat att="syntacticConstituent" val="NP"/>
    </SyntacticArgument>
</SubcategorizationFrame>
```
The multilingual Extension of LMF

Example for consensual standardization
The multilingual extension of LMF

Three approaches to model contrastive information

- Interlingua:
  - Reference to “Sense Axis”, as in Papillion (http://www.papillon-dictionary.org)
  - Reference to external content (“External Reference”), e.g. ontology

- Transfer
  - Tests for source and target items
  - make reference to syntactic properties (of source)

- Translation Memory Systems, EBMT:
  - List contexts providing translation information
A Dictionary Entry

clergyman, common noun, singular, pl: clergymen

Feature Structure Representation:

- **partOfSpeech**: common noun
- **Lemma**: written: 1 clergyman
- **Wordform**: writtenForm 1
  - **grammaticalNumber**: singular
- **Wordform**: writtenForm clergymen
  - **grammaticalNumber**: plural

LMF language

- 1 Lexical Entry
- 1 Lemma
- 2 WordForm
- extensions of any kind
XML instance of LMF

```xml
<?xml version="1.0" encoding="UTF-8"?>
<LexicalResource>
    <GlobalInformation>
        <feat att="languageCoding" val="ISO 639-3"/>
    </GlobalInformation>
    <Lexicon>
        <feat att="language" val="eng"/>
        <LexicalEntry>
            <feat att="partOfSpeech" val="commonNoun"/>
            <Lemma>
                <feat att="writtenForm" val="clergyman"/>
            </Lemma>
            <WordForm>
                <feat att="writtenForm" val="clergyman"/>
                <feat att="grammaticalNumber" val="singular"/>
            </WordForm>
            <WordForm>
                <feat att="writtenForm" val="clergymen"/>
                <feat att="grammaticalNumber" val="plural"/>
            </WordForm>
        </LexicalEntry>
    </Lexicon>
</LexicalResource>
```
Another XML instance of LMF

<Lexicon>
  <Lexiconentry>
    <lemma>clergyman</lemma>
    <form type="sgl">clergyman</form>
    <relatedform>clergymen</relatedform>
    <pos>CNoun</pos>
  </Lexiconentry>
  <Lexiconentry>
    <lemma>clergymen</lemma>
    <form type="pl">clergymen</form>
    <relatedform>clERGYman</relatedform>
    <pos>CNoun</pos>
  </Lexiconentry>
</Lexicon>
Another instance of LMF

Clergyman:
<partOfSpeech> == commonNoun
<lemma> == clergyman
<writtenform sgl> == clergyman
<writtenform pl> == clergymen
Outline

Objectives of LMF

LMF Philosophy

The LMF ‘Core Package’

LMF Extensions

Benefit of LMF
Usability of LMF

- common meta language
- Instantiation not determined = no standardization (!)
  - XML
  - DATR
  - CSV
  - FSR

Mapping possible?

- important: we need precise knowledge what features stand for