Modeling full form lexica for Arabic

Susanne Alt
Amine Akrout
_Atilf-CNRS_

Laurent Romary
_Loria-CNRS_
Objectives

- Presentation of the current standardization activity in the domain of lexical data modeling
- Validation of the proposed standard on Arabic
- Contribution to the establishment of a reference resource for Arabic
Overview

- Background
  - Why do we need full form lexica?
- Standards
  - Lexical resources & dictionaries
- Instanciation
  - Specificities of an Arabic full form lexicon
- Overall goal
  - making current work interoperable
Two views on lexical data

- **Extensional representation**
  - exhaustive list of observables
    - set of inflected word forms
    - set of syntactic constructions

- **Intensional representation**
  - factorization of regular behaviour ("grammar")
    - lemma + inflection rules
    - deep syntactic representation + transformation rules
Full form lexica: advantages

- Local linguistic information
  - local inflectional variants (*courbattu vs. courbaturé*)
  - defective paradigms (*nous pleuvons*)
  - phonological variants (*les – [le]/[lez]*)

- Testimony of inflected forms
  - token frequency wrt a reference corpus

- Exchange of lexical resources
  - no consensus on encoding format for grammar rules
  - pivot format for merging and comparing lexicons

- Extensional data for data recognition purposes
Standards for NLP lexica

- Forefathers
  - a wide range of international projects
    - MULTEXT, EAGLES, ISLE/MILE, Parole...

- XML encoding of print dictionaries
  - "Print dictionary" chapter of the TEI
    - http://www.tei-c.org

- Terminology
  - Sense-to-word oriented
    - Terminology Markup Framework (ISO 16642)
Lexical Markup Framework

- Future ISO standard 24613
- ISO technical committee TC 37/SC 4
  - Language Resource Management
    - [http://www.tc37sc4.org](http://www.tc37sc4.org)
    - [http://lirics.loria.fr](http://lirics.loria.fr)
- Project leaders
  - Monte George (USA) & Gil Francopoulo (FR)
- First applications
  - *Morphalou* (Salmon-Alt et alii, 2004)
LMF: Basic principles

- An open platform for specifying lexical data
  - implemented prototypes: Lexus, Syntax
- Main modeling principles
  - metamodel
    - basic building blocks and basic structural constraints
    - e.g. "A lexical database is made of lexical entries."
  - data categories
    - basic linguistic descriptors
    - e.g. "grammatical gender", "synonymOf", ...
    - stored in a shared data category registry
LMF core metamodel

Lexical Database

Global Information

Lexical Entry

Form

Sense
Data categories

- Independent from the hierarchical structure of the data model
  - /partOfSpeech/, /grammaticalNumber/, /grammaticalCase/

- Characteristics
  - complex vs. simple
    - /grammaticalNumber/ => /singular/, /plural/
  - relational data categories
    - /synonymOf/, /toInflectionalParadigm/
  - generic vs. language specific
    - /grammaticalNumber/ => {/singular/, /plural/, /dual/}
**Entry Identifier:** /grammaticalGender/

**Profile:** Morpho-syntax

**Definition:** Grammatical genders are classes of nouns reflected in the behavior of associated words

**Explanation:** Grammatical gender is distinguished from natural gender by the fact that grammatical gender requires *agreement* between nouns and the forms of modifiers ...


**Range:** 

<table>
<thead>
<tr>
<th>Object Language</th>
<th>Name</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr</td>
<td>genre</td>
<td>{masculine/, feminine/}</td>
</tr>
<tr>
<td>en</td>
<td>gender, grammatical gender</td>
<td>{}</td>
</tr>
<tr>
<td>de</td>
<td>Genus, Geschlecht</td>
<td>{masculine/, feminine/, neuter/}</td>
</tr>
</tbody>
</table>

---

### Object Language: fr
- **Name:** genre
- **Range:** {masculine/, feminine/}

### Object Language: en
- **Name:** gender, grammatical gender
- **Range:** {}

### Object Language: de
- **Name:** Genus, Geschlecht
- **Range:** {masculine/, feminine/, neuter/}

---

**Documentation and localization**
Lexicon specification

Lexicon Database
  ├── Global Information
  │    └── Form
  └── Lexical Entry
      └── Sense

/grammaticalCategory/
GMT (Generic Mapping Tool)

<struct type="lexicalDatabase">
  <struct type="globalInformation">...
  <struct type="lexicalEntry">
    <feat type="grammaticalCategory">...
    <struct type="form">...
    <struct type="sense">...
    <struct type="sense">...
    ...
  </struct>
  </struct>
</struct>
User specific XML format

<lexicalDatabase>
  <globalInformation>...</globalInformation>
  <lexicalEntry POS="...">
    <form>...</form>
    <sense>...</sense>
  </lexicalEntry>
  <lexicalEntry POS="...">
    <form>...</form>
    <sense>...</sense>
  </lexicalEntry>
  ...
</lexicalDatabase>
Applying LMF to Arabic

- Little representation of Arabic speaking countries in ISO/TC 37/SC 4
- NLP of Arabic morphology
  - Beesley K., 2001; Buckwalter, 2002; Cavalli-Sforza et alii, 2000; Maamouri & Bies, 2004; Tahir et alii, 2004…
- Yet, no widely, freely accessible and cumulative lexicon can be used to boost research on Arabic language
  - strategy: combining efforts through standardization
FR vs. Arabic full form lexica

- French lexicography
  - semasiological + alphabetical perspective
- (Traditional) Arabic perspective
  - mixed + root based
    - grouping of all derivates from consonant pattern
      - ktb (notion of writing) => kātaba (to write), kattaba (cause to write), maktabun (desk), maktabatun (library), kitābun (book)
  - therefore
    - distinction between human readability and machine processing
    - essential to keep reference to the root
Adapting LMF to Arabic (I)

- Specifying the notion of "lexical entry"
  - alphabetically ordered
  - characterized by
    - POS
    - keyform
    - reference to the root
Adapting LMF to Arabic (II)

- Specifying the notion of "inflected form"
  - a word form and inflectional features
  - form related & inflection related data categories

```
Inflected Form

/orthography/
/pronunciation/
/grammaticalGender/
/grammaticalNumber/
/grammaticalCase/
/grammaticalDefiniteness/
/grammaticalAspect/
/grammaticalVoice/
/grammaticalMood/
/grammaticalPerson/
```
Adapting LMF to Arabic (III)

- Form related data categories
  - orthography and pronunciation
  - both are subject to refinements ("local metadata")
    - transliteration: fully reversible one-to-one mapping to original orthography
      - Buckwalter transliteration
    - transcription: devised to render (morpho)phonology
      - IPA

/orthography/ => /transliteration/
/pronunciation/ => /transcription/
Adapting LMF to Arabic (IV)

- Some questions on inflection related data categories
  - Nouns
    - `/grammaticalGender/` => `/masculine/`, `/feminine/`
      - no lexicalized (because of gender change in plural forms)
      - choice of no "underspecified" gender
    - `/grammaticalNumber/` => `/singular/`, `/plural/`, `/dual/`
      - (enter) and/or pick up `/dual/` from the DCR
    - `/grammaticalCase/` => `/nominative/`, `/accusative/`, `/prepositional/`
      - terminology (prepositional, indirect, possessive or genitive)?
    - `/definiteness/` => `/definite/`, `/indefinite/`
      - one or two categories of definiteness (def. alkitâbu, pos. kitâbî)?
  - inflection vs composition (e.g. prepositional affixes)?
The fully specified model

Lexical Database

Global Information

Lexical Entry

Word Form Set

Inflected Form

Form

/orthography/, /pronunciation/

Inflection

/grammaticalCategory/
/keyForm/
/gloss/
/root/

/grammaticalGender/
/grammaticalNumber/
/grammaticalCase/
/grammaticalDefiniteness/
/grammaticalAspect/
/grammaticalVoice/
/grammaticalMood/
/grammaticalPerson/
<lexicalEntry keyform="kataba" grammaticalCategory="verb" root="ktb" gloss="écrire">
  <wordFormSet>
    <inflectedForm>
      <form>
        <orthography code=""Akrout_2005">katabtu</orthography>
        <realization/>
      </form>
      <inflection>
        <grammaticalAspect>perfect</grammaticalAspect>
        <grammaticalGender>masculine</grammaticalGender>
        <grammaticalPerson>firstPerson</grammaticalPerson>
        <grammaticalNumber>singular</grammaticalNumber>
        <grammaticalVoice>active</grammaticalVoice>
      </inflection>
    </inflectedForm>
    ...
    <inflectedForm>
      <form>
        <orthography code="Akrout_2005">taktubâ</orthography>
        <realization/>
      </form>
      <inflection>
        <grammaticalAspect>imperfect</grammaticalAspect>
        <grammaticalGender>masculine</grammaticalGender>
        <grammaticalPerson>secondPerson</grammaticalPerson>
        <grammaticalNumber>dual</grammaticalNumber>
        <grammaticalMood>subjunctive</grammaticalMood>
      </inflection>
    </inflectedForm>
    ...
  </wordFormSet>
</lexicalEntry>
Towards a reference lexicon for Arabic: issues

- Interoperability
  - Comparison of proprietary specifications

- Coverage
  - Completion of specific advances (dialectal, terminological, phonology)

- Accessibility
  - Common query interface, wide (free ?) distribution

- Maintenance
  - Common rules to ensure editorial evenness
  - Documentation & user manuals

- A step towards an intensional representation…