The place of lexicography in (computer) science
Laurent Romary

To cite this version:
Laurent Romary. The place of lexicography in (computer) science. The Future of Academic Lexicography: Linguistic Knowledge Codification in the Era of Big Data and AI, Frieda Steurs; Dirk Geeraerts; Niels Schiller; Marian Klamer; Iztok Kosem, Nov 2019, Leiden, Netherlands. hal-02358218
The place of lexicography in (computer) science

Laurent Romary, Inria, team ALMAnaCH
Overview

Understanding the role of lexicographic work in other scholarly fields:

• Dictionaries as primary sources in the humanities

• The CS perspective: from data modelling to data mining
  • Lexicography as a rich playground for data modelling
    • Current developments in international standardisation
  • Automatic analysis of legacy print dictionaries
    • The GROBID-dictionary experience

• Perspectives – automatic data enrichments
Lexicographic works as a primary source

• Dictionaries integrate a wealth of linguistic information, but also represent a mirror of their times
  • The objective of comprehensiveness makes them essential primary sources of further humanities studies

• A wide range of possible reuse possibilities, illustrated through 2 examples:
  • Digitising the *Dictionnaire Universel* from Trévoux-Basnage (ANR project BASNUM)
  • The *Vocabulario en lengva misteca* at the service a language documentation project
An encyclopaedic witness – digitising the 1701 Dictionnaire Universel

• The *Dictionnaire Universel* (*DU*), the first truly encyclopaedic dictionary
  • Covers general language, but above all terms from arts, crafts and sciences
  • Highly influential throughout Europe both directly and indirectly

• Antoine Furetière (1619-1688)
  • Ex-member of the *Académie française*, at loggerheads over his personal universal dictionary. The *DU* was published posthumously in 1690 in the Netherlands
  • Created the *Dictionnaire universel* as an encyclopaedic dictionary including all words used in France of his day

• Henri Basnagé de Beauval (1657-1710)
  • A Protestant lawyer, son of a leading member of the Parliament of Normandy
  • Forced into exile after the Revocation of the Edict of Nantes that outlawed protestants
  • Succeeds the protestant philosopher Pierre Bayle as as editor of a literary and philosophical journal - *Histoire des ouvrages des savants*
  • Engaged by Leers, publisher of the *DU*, to compile a revised and enlarged version, published 1701
  • Uses experts to write scientific entries

ANR project BASNUM
Colleagues involved: Geoffrey Williams, Mohamed Khemakhem (Univ. de Grenoble), Ioana Galleron, Clarissa Stincone (Univ. Sorbonne Nouvelle), Benoit Sagot, Laurent Romary, Pedro Ortiz (Inria)
The complex editorial history of the *Dictionnaire universel*

1690: first DU version, written and published by Antoine Furetière
   • normative approach

1701: *second DU version, entirely revised by Basnage de Beauval and much augmented (1/3)*
   • *descriptive approach*

1702: reprint of Basnage’s version in 2 volumes

1708: second reprint (3 vols)

1725: new version (4 vols) revised by Brutel de La Rivière
CARME. f. m. Ordre de Religieux, qui est l'un des quatre Mendians, qui pretend tirer fon nom du Mont Carmel en Syrie, qu'on dit avoir ete habite par Elie. Ils ont ete amenez en France par le Roi Louis IX. Il est celebre par la devotion du Scapulaire, & par la vision de Simon Stock, a qui il fut donné par la Sainte Vierge. Sur quoy de Launoy a ecrit une curieuse Dissertation. La Vierge attache ce privilege au Scapulaire, & l'habit des Carmes, que ceux qui meurent le Samedi chargen de ces pieues depouilles, sont exemptes des flames du Purgatoire. Les Carmes se difent oncles de J. Christ, & freres de la Vierge. On dit les Carmes du grand Couvent; les Carmes Mitigez, qu'on nommee a Paris Billettes; & les Carmes Dechaussiez, qui ont ete reformes des autres. Dans des Theses soutenues a Beziers mentionnees dans le Journal de Hollande, on dit qu'il est fort probable que Pythagore estoit Carme, & que les Druides des Gaulois avoient aussi les observances regulieres des Carmes.

MONT CARMEL, est un Ordre Militaire de Chevaliers Hospitalliers, fondé par le Roi Henri IV, sous le titre, l'habit & la Regle de Notre Dame du Mont Carmel; & en consequence des Bulles du 16. Fevrier 1607, il a été uni a l'Ordre des Chevaliers de St. Lazare de Jerusalem, par acte du dernier Octobre 1608, avec toutes les Commenderies, Prierez & autres bien pour la dotation.

Carme, est aussi un epee de dacier. Voyez Acier.
Carme, est aussi un vieux mot qui signifiait un vers. Il vient du Latin carmen; & en ce sens il est tout-à-fait hors d'usage.
Understanding the dictionary: some humanities research questions

• What source texts? Dictionaries and other texts?
• What links between the DU and *Histoire des ouvrages des savants*
• What lexicographical model between prescription and description?
  ➢ the place of ‘good’ usage and the ‘best’ authors
  ➢ the dictionary as a language teaching tool – what users? what means?
  ➢ The characterisation of terms
  ➢ The role of contemporary scientific and literary networks
• To what degree the 1701 DU was theologically a "protestant dictionary”
• What changes were brought in between 1690 and 1701, between 1701 and 1725/27, and between 1701 and the 1704 Trévoux
• Who authored which entries in 1701? - Basnage and his specialist informers.
Linguistic description of Mixtepec-Mixtec

- Sa’an Savi “rain language”
- ISO 639-3 code: ‘mix’
- San Juan de Mixtepec - Juxtlahuaca district (Oaxaca, MEX)
- “Vigorous” status but highly under-resourced
- Oto-Manguean, Mixtecan, Mixtec-Cuicatec, Mixtepec-Mixtec
- Tonal
- Spoken data mostly collected in sessions working with speakers from a small village called Yucunani in the San Juan Mixtepec municipality
- Estimated (+-9,000 -10,000 speakers)
  
  \textit{Source: (INEGI, 2010)}
- Phonology has been studied by Pike and Ibach (1978); Paster and Beam de Azcona (2004-2007);
- Beckman and Nieves-SIL (2005-current) published booklets and are working on developing orthography
The research project

• Language documentation: (Jack Bowers’ PhD)

• Primary sources of language data:
  • Speaker consultations (recordings, new written material..)
  • +- 40 Children’s Booklets (SIL)
  • Public sources (YouTube, other)
  • Examples from academic papers

• Goals:
  • TEI Corpus
  • Linguistic descriptions
  • TEI Dictionary (actually 2 dictionaries, 1 general, 1 inflectional)
  • (Etymology) would like to create data contents and structure that can be copied and integrated into treatment of related languages

Jack Bowers, Laurent Romary. TEI and the Mixtepec-Mixtec corpus: data integration, annotation and normalization of heterogeneous data for an under-resourced language. 6th International Conference on Language Documentation and Conservation (ICLDC), Feb 2019, Honolulu, United States. ⟨hal-02075475⟩
Overview of the Source & Output

• ‘Vocabulario en lengva misteca’ published by the Dominican Francisco de Alvarado (1593)

• Variety from Teposcolula Mexico (Mixteca Alta)
  • Classical Mixtec/Colonial Mixtec/Yucu Ndaa

• Entries based on three earlier dictionary sources:
  • Castilian-Nahuatl (Valley of Mexico, 1571)
  • Castilian-Zapotec (Valley of Oaxaca, 1578)
  • Castilian-Latin (1492)

• PDF re-organized, modernized version ‘Voces de Dzaha Dzahui’ (Jansen & Pérez Jiménez, 2009)

• TEI dictionary produced contains roughly 26,600 entries and related entries.
Versions of the resource

• Original (Printed: 1593) > (facsimile edition 1965)
• Mesolore (Bakewell & Hamman, 2001)
• Digitized from scanned copy
• Jansen and Pérez Jimenez 2009
Utility/Purpose of Endeavour

• Increase coverage of relevant lexical material in Mixtepec-Mixtec documentation (ISO 639-3 [mix])
  • Link and cross-reference in Mixtepec TEI dictionary
• Machine searchable data set for:
  • Study of the Yucu Ndaa variety
  • Historiographical and philological research
• Create a more cohesive body of pan-Mixtecan resources
  • Vocabulary for cross Mixtecan comparison; (81 Varieties of Mixtec)
• TEI format can easily be exported into other formats for non-TEI users
Integration into Mixtepec-Mixtec Project: TEI Structure of Output

- Goal to match the structure used in the Mixtepec-Mixtec TEI dictionary (Bowers & Romary 2018)

```
<entry xml:id="fruit-plantain">
  <form type="lemma">
    <orth xml:lang="mix">nchika</orth>
    <pron xml:lang="mix" notation="ipa">njiká</pron>
  </form>
  <gramGrp>
    <pos>noun</pos>
  </gramGrp>
  <sense corresp="http://dbpedia.org/resource/Plantain">
    <usg type="domain">Fruit</usg>
    <cit type="translation">
      <form>
        <orth xml:lang="en">plantain</orth>
      </form>
    </cit>
    <cit type="translation">
      <form>
        <orth xml:lang="es">plátano</orth>
      </form>
    </cit>
  </sense>
</entry>
```

```
<entry xml:id="plátano">
  <form type="lemma">
    <orth>chita</orth>
  </form>
  <gramGrp>
    <pos>noun</pos>
  </gramGrp>
  <sense corresp="http://dbpedia.org/resource/Plantain">
    <usg type="domain">Fruit</usg>
    <def xml:lang="es">plátano</def>
    <def xml:lang="en">plantain</def>
  </sense>
</entry>
```

**nchika** [ŋʒiká] (noun) [FRUIT] plantain, plátano

**chita** (noun) [FRUIT] plantain, plátano

Classical Mixtec

Mixtepec-Mixtec
Going further: modelling and standardising lexical resources
Lexical resources in their varieties

• A variety of contexts and forms
  • Legacy dictionaries, dialectological studies, NLP lexica
  • Full form, etymology, corpus based research
  • Word document, database, shoebox, proprietary XML...
  • Lexical vs. Editorial views
  • Onomasiological vs. semasiological structures
Lexicography or terminology

• Lexicography
  • Generic view on “words”
  • Attempt to provide a large coverage of a language
  • Semasiological view
    • Word > meaning(s)

• Terminology
  • Term: form associated to a specific concept within a given domain
  • Onomasiological view
    • Concept > various possible linguistic forms

• Depends on available data, objectives and user scenarios
Comparing approaches

Semasiological approach
• Large coverage
• All parts of speech
• Build-in polysemy
  • Multiple senses for the same entry
• Referential synonymy

Onomasiological approach
• Domain oriented
• Essentially nouns
  • Extension to verbs, adjectives
• No polysemy (needs to be reconstructed)
• Build-in synonymy
  • Multiple terms for the same concept
Basic modelling of lexical components

Semasiological models

- Lexical entry
  - Form
    - Form-related descriptions
  - Sense
    - Sense-related descriptions

Issues

- Various levels/sensibilities in entry groupings: homonyms, families (e.g. roots)
- Providing a neat way of representing lexical dependencies: from “see also” to multi-word expressions

Onomasiological models

- Terminological entry
  - Concept-related descriptions
  - Language-related descriptions
  - Term-related descriptions

Issues

- Representing conceptual relations between entries
- Providing fine-grained semantic information at term level (e.g. usage, translation equivalents)

Issues

- Making the appropriate choice of model
- Integrating information between the two types of models
Why standardizing all this?

• Defining methods, models and format to facilitate
  • Exchange of lexical data
  • Pooling heterogeneous lexical data
  • Interoperability between software components
    • Search engines, layout, extraction of linguistic properties
• Comparability of results
  • E.g. Linguistic coverage of lexical databases
  • Exchange of ideas within a community with a common background
Standardization initiatives for lexical/terminological resources

- **TEI**
  - P5 edition of the guidelines
    - Cf. specification platform (ODD)
    - Dictionary chapter
    - Former terminology chapter (ancestor of TBX)

- **ISO**
  - ISO/TC 37: Language and terminology
    - ISO/TC 37/SC 3: ISO 16642 (TMF), ISO 30042 (TBX)
    - ISO/TC 37/SC 4: ISO 24613 (LMF)

- **W3C**
  - SKOS, Ontolex
In the beginning

Text archives
Humanities Standards
SGML

Not intended (immediately) for individual scholars

1. Novembre 1987: Vassar College, Poughkeepsie
A quick historical overview

• 1960’s — GML (Generalized Markup Language) by IBM
• 1970’s & 1980’s — ANSI initiates project to develop a Standard text-description language based on GML
• 1983 — SGML becomes an industry standard
• **1987 — TEI (Text Encoding Initiative)**
• 1990 — HTML 1.0 (HyperText Markup Language)
• 1992 — TEI edition P3 (Michael Sperberg-McQueen and Lou Burnard, eds)
• 1997/1998 — XML 1.0 (eXtensible Markup Language) (Tim Bray, Jean Paoli and Michael Sperberg-McQueen, eds)
The TEI Dictionary chapter

• Initially designed within a working group lead by N. Ide and J. Veronis
• Accounts for both presentational and database views
  • Cf. <entry>, <entryFree>, ... and <dictScrap>
• Based on a hierarchical abstract model (crystals)
  • <form>: for characterizing the orthographic or phonetic form of the word
    • <orth>, <pron>, etc.
  • <gramGrp>: grammatical features
    • May characterize an entry, a specific form or a specific sense
    • <pos>, <gen>, generic <gram> feature
• <sense>: iterative and recursive
  • May contains definitions, examples, etymological information, translations, etc.
IRL: Petit Larousse illustré (1906)

<entry xml:id="pléthore" n="1906-001_unknown">
  <form type="lemma"><orth>PLÉTHORE</orth></form>
  <gramGrp><pos expand="nom">n.</pos>
    <gen expand="féminin">f.</gen></gramGrp>
  <etym><pc>(</pc>du <lang expand="grec">gr.</lang>
    <mentioned>plêthorê</mentioned><pc>,</pc>
    <gloss>plénitude</gloss></pc>)</pc><pc>.</pc></etym>
  <sense><def>Surabondance de sang, d'humeurs</def><pc>.</pc></sense>
  <sense><usg type="style" rend="italic" expand="figuré">Fig.</usg>
    <def>Surabondance quelconque amenant un état fâcheux</def>
    <pc>:</pc>
    <cit type="example">
      <quote>la pléthore des capitaux cause la diminution du taux de l'intérêt</quote>
    </cit></pc></sense>
</entry>

Advantages of being in the TEI framework

- Benefitting from the TEI environment
  - TEI Modelling Language: ODD
    - Customizing the guidelines within a project (e.g. restraining possible values)
  - Availability of a wealth of additional elements (~600)
    - E.g. annotating textual content, reflecting the specificities of the source etc.

- Standardisation reactivity
  - Issuing GitHub tickets for resolving bugs or introducing new features

- A community of experts
  - Support through the mailing list

- Main characteristic (drawback?): +very+ flexible
Going ISO to provide a stable background

• Advantages of going ISO
  • International approval of ISO members
    • And international expert participation by construction
  • Stable background that is easy to reference (and known by third parties, non-linguistic geeks, our institutions etc.)

• From a lexical point of view
  • Providing a generic model, independently of any specific implementation/serialisation
    • Stabilizing concepts, constraints and vocabulary
  • With the on-going LMF revision: Introducing the TEI as one possible model
LMF

ISO 24613:2008 Language resource management — Lexical markup framework (LMF)

• Developed within ISO TC37/SC4/WG4
  • TC 37: Language and terminology
  • TC37/SC4/WG4: Lexical resources

• Shortcomings
  • Bulky: a single document with annexes
    • Major hindrance to revision
  • Complex modelling
    • Complex relationships between classes, redundant mapping mechanisms
  • Complex and ad hoc serialisation
  • Does not cover prominent information: Etymology and Diachrony
LMF Reloaded: Abstract Modelling

Restructuring: Multi-part standard
• ISO 24613-1 - Core model (published in June 2019)
• ISO 24613-2 - Machine Readable Dictionaries (MRD) model
• ISO 24613-3 - Diachrony-Etymology
• ISO 24613-4 - TEI serialisation
• ISO 24613-5 - LBX serialisation
• ISO 24613-6 - Syntax and Semantics
• ISO 24613-7 - Morphology
Language resource management — Lexical markup framework (LMF) — Part 1: Core model
From LMF to TEI – serialising one with the other (Part 4 – TEI serialisation)

Background: Laurent Romary. TEI and LMF crosswalks. *JLCL - Journal for Language Technology and Computational Linguistics*, 2015, 30 (1). ([hal-00762664v4](hal-00762664v4))
Example: inflectional (full-form) lexicon
LMF reloaded - Etymology

ISO 24613-3, prepared for DIS ballot (as of November 2019)
new classes: Etymology, Etymon, Cognate and EtyLink

Derivatives: center, centre, intr. and tr. v., center-ing, centr-ing, centre-ing, n.

Source: Klein’s Comprehensive dictionary of the English Language
TEI lex 0: tightening the guidelines

• Initiative set up in the context of the DARIAH working group on lexical resources, supported by the EU project Elexis

• Objectives
  • Designing a target format for heterogeneous lexical data integration

• Trade-off
  • Compliance with the standard
  • Fine-tuning to the needs of a specific context/scenario

• Back to what a standard is: a common reference for a transaction
  • Perfect to say: “I am compliant to standard X except for...”
  • The TEI guidelines provides the means to carry out such customizations
Application – the ELEXIS project

• European Lexicographic Infrastructure
  • 2018-02-01 – 2022-01-31
  • “integrate, extend and harmonise national and regional efforts in the field of lexicography”
    • Focuses on “efficient access”
  • Cooperation with CLARIN and DARIAH for long-term sustainability

• Lexical formats and standards in ELEXIS
  • Double sided approach TEI – Ontolex
  • TEI Lex 0 specification at the core of the data hub
The ELEXIS centralized hub

Dict 1
Dict 2
Dict...
Dict n

TEI Lex 0 “soup”

ODD

Ontolex delivery

User
Enforcing the semasiological model

<entry>
  <form type="lemma">
    ...
  </form>
  <sense>
    <def>...</def>
  </sense>
</entry>
Simplifying the dictionary micro-structure

- **Current situation**
  - Containing vs. contained entries
    - `<superEntry>` – `<entry>` – `<re>`
  - Structured vs. unstructured entries
    - `<entry>` – `<entryFree>`

- **The TEI Lex-0 vision**
  - Representing all entry-like objects as `<entry>`
    - Making `<entry>` recursive
    - Making more use of `<dictScrap>`
Recursive entry
- example

<entry type="wordFamily">
  <form type="base">
    <orth>Haus-</orth>
  </form>
  <pc>.,</pc>
  <form type="base">
    <orth>haus-</orth>
  </form>
  <pc>:</pc>
  <!-- possibly some shared usage information -->
</entry>

<entry type="wordForm">
  <form type="lemma">
    <orth expand="Hausaltar">-altar</orth>
    <pc>.,</pc>
    <gramGrp>
      <gen value="masculine">der</gen>
    </gramGrp>
  </form>
  <sense>...</sense>
</entry>

<entry type="wordForm">
  <form type="lemma">
    <orth expand="Hausandacht">-andacht</orth>
    <pc>.,</pc>
  </form>
  <!-- ...
   -->
</entry>

<entry>...
</entry>
Do we really have to encode all this manually?
Considering machine learning techniques

- Strong layout regularities within a given dictionary
- Similarities within a family of dictionaries
- Supervised – unsupervised $\iff$ Less data – more data
- Features to be considered:
  - Layout
  - Lexical
- Sequencing task: well adapted for so-called graphical models
Why GROBID?

“Converting PDF to XML is a bit like converting hamburgers into cows. You may be best off printing it and then scanning the result through a decent OCR package.” Michael Kay (http://lists.xml.org/archives/xml-dev/200607/msg00509.html)

Inspired from: Duncan Hull
GROBID-Dictionaries

• Automatic extraction of TEI structures from digitised dictionaries (Khemakhem et al. 2017)
  • Input: PDF (soon ALTO)
  • Output: TEI compliant lexical resource

• Spin-off from GROBID (Romary and Lopez 2015)
  • Initiated in 2007
  • Automatic extraction of structural data from scholarly papers
    • Metadata (author, title, affiliations, keywords, abstract), bibliography, ... full text
  • And open source...

• Uses Conditional Random Fields (CRF) (Lavergne et al. 2010)
  • Probabilistic models for sequence labelling tasks
State Of the Art

• Rule based approaches dominate

• Few machine learning attempts
  • Promote CRF for sequence labelling in dictionaries (Crist 2011)
  • Reduce the annotation time for labels (Bago et al. 2015)
Cascading CRF models
Cascading CRF models

https://github.com/MedKhem/grobid-dictionaries
First experiments

Easier English Basic Dictionary (EEBD)

Fang - French Dictionary (FFD)
Evaluation: Token Level - F1 Score

- **Lexical Entry**
  - EEBD: 100 LE (8 pages): 76 training, 24 evaluation
  - FFD: 71 LE (3 pages): 47 training, 24 evaluation

- **Sense**
  - EEBD: 30 blocks (6 pages): 15 training, 15 evaluation
  - FFD: 90 LE (4 pages): 71 training, 19 evaluation
Further complexities
## Experiments: Lexical Entry Model

<table>
<thead>
<tr>
<th>TEI element</th>
<th>Sample 1</th>
<th></th>
<th></th>
<th>Sample 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td>&lt;etym&gt;</td>
<td>87.5</td>
<td>60</td>
<td>71.19</td>
<td>73.68</td>
<td>71.79</td>
<td>72.73</td>
</tr>
<tr>
<td>&lt;form&gt;</td>
<td>94.44</td>
<td>92.73</td>
<td>93.58</td>
<td>92.24</td>
<td>96.4</td>
<td>94.27</td>
</tr>
<tr>
<td>&lt;pc&gt;</td>
<td>90.91</td>
<td>69.44</td>
<td>78.74</td>
<td>88.97</td>
<td>80.13</td>
<td>84.32</td>
</tr>
<tr>
<td>&lt;re&gt;</td>
<td>33.33</td>
<td>9.09</td>
<td>14.29</td>
<td>55.56</td>
<td>22.73</td>
<td>32.26</td>
</tr>
<tr>
<td>&lt;sense&gt;</td>
<td>67.65</td>
<td>59.28</td>
<td>63.19</td>
<td>77</td>
<td>76.65</td>
<td>76.84</td>
</tr>
<tr>
<td>&lt;xr&gt;</td>
<td>100</td>
<td>80</td>
<td>88.89</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 3: Field Level Evaluation of the Lexical Entry Model
GROBID: the next generation

Pedro Javier Ortiz Suárez, Laurent Romary, Benoît Sagot. Preparing the Dictionnaire Universel for Automatic Enrichment. 10th International Conference on Historical Lexicography and Lexicology (ICHLL), Jun 2019, Leeuwarden, Netherlands. [hal-02131598]
Wrapping up

• Dictionaries are cool things
  • But we all share this...

• A great deal of standardisation work has already been done
  • A strong basis for improving interoperability
  • Further convergence work is needed

• Huge expectation around automatic annotation
  • Cf. Basnage: a drop in the legacy ocean (e.g. SIL)
    • But lack of generalisation across dictionaries
  • Future:
    • Families of dictionaries
    • Deep learning

• Towards a wealth of dictionary sources
  • Changing scale for more lexical knowledge
  • Towards a lexical time-space machine => bringing back knowledge to the lexicographic folk
Merci pour votre attention!