Multilingual content management and standards with a view on AI developments

Laurent Romary

Multilingual content management and standards with a view on AI developments

Laurent Romary

Directeur de Recherche, Inria, team ALMAnaCH

ISO TC 37, chair

Language and AI

- Central role of language in the revival of AI (machine-learning based models)
  - Applications: document management and understanding, chatbots, machine translation
  - Information sources: public (web, cultural heritage repositories) and private (Siri, Amazon Alexa) linguistic information
  - European context: cf. *Europe's Languages in the Digital Age*, META-NET White Paper Series
- Variety of linguistic forms
  - Spoken, written, chats and forums
  - Multilingualism, accents, dialects, technical domains, registers, language learners
    - General notion of language variety
- Classifying and referencing the relevant features
  - Role of standards and standards developing organization (SDO)

A concrete example for a start

Large scale corpus

OSCAR

BERT

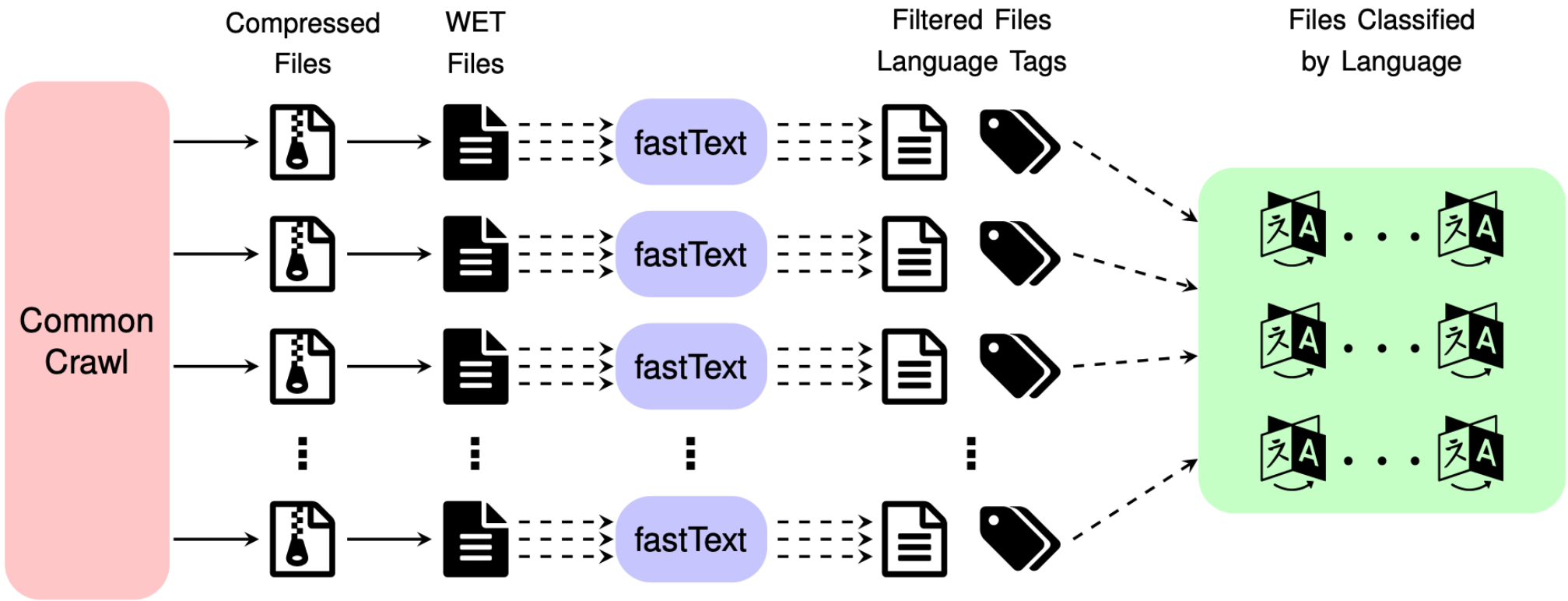
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Language model

CamemBERT

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, et al.. CamemBERT: a Tasty French Language Model. *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*, Jul 2020, Seattle / Virtual, United States. ⟨10.18653/v1/2020.acl-main.645⟩. ⟨hal-02889805⟩

The infrastructure to create OSCAR - goclassy



Pedro Javier Ortiz Suárez, Benoît Sagot, Laurent Romary. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Jul 2019, Cardiff, United Kingdom. ⟨10.14618/IDS-PUB-9021⟩. ⟨hal-02148693⟩

OSCAR

Language	Size		Words	
	Orig	Dedup	Orig	Dedup
English	2.3T	1.2T	418,187,793,408	215,841,256,971
Russian	1.2T	568G	92,522,407,837	46,692,691,520
Spanish	278G	149G	47,545,122,279	25,928,290,729
French	282G	138G	46,896,036,417	23,206,776,649
German	308G	145G	44,878,908,446	21,529,164,172
Italian	137G	69G	22,248,707,341	11,250,012,896
Portuguese	124G	64G	20,641,903,898	10,751,156,918
Chinese	508G	249G	14,986,424,850	6,350,215,113
Japanese	216G	106G	4,962,979,182	1,123,067,063
Polish	109G	47G	15,277,255,137	6,708,709,674
Total OSCAR	6.3T	3.2T	844,315,434,723	425,651,344,234

Language and standards – why?

- A general issue of reproducibility
  - Can we compare two linguistic descriptions
    - E.g. two etymological assumptions in a dictionary entry
- An acuter context created by the machine learning revival
  - Comparing results of AI components
    - Issues related to segmentation, categorisation (noun, verbs, feminine, dual, etc.), dependencies towards other linguistic descriptions (lexica, lower linguistic levels)
  - Reproducibility of results
    - Can we compare two training campaigns?
    - How do we deal with cross linguistic comparison?
    - Precise qualification (e.g. coverage) and identification of training corpora

Standards

- What they are
  - Reference background for the management of a technical process
  - The three pillars of a standard: consensus, publicity, maintenance
  - A constraint for each, a benefit for all!
    - Reading it, understanding it, implementing it…
- What they aren't
  - Something coming from nowhere… participating is always an option
  - Mandatory documents: rather a common ground for a transaction
  - Regulations: unless a state or an international organisation takes it up
- The need for standards developing organizations (SDO)
  - Ensure that the three pillars of standardisation are in place
- Standards are like the EU: you need to understand how they work to appreciate them…

Standards and AI

- A new committee has been set up
  - IT context: ISO/IEC JTC 1/SC 42 Artificial intelligence
  - Work program - 3 publications so far:
    - ISO/IEC 20546:2019 Information technology — Big data — Overview and vocabulary
    - ISO/IEC TR 20547-2:2018 Information technology — Big data reference architecture — Part 2: Use cases and derived requirements
    - ISO/IEC TR 20547-5:2018 Information technology — Big data reference architecture — Part 5: Standards roadmap
  - Numerous ongoing activities:
    - Concepts and terminology, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML), Risk Management, Assessment of the robustness of neural networks, Overview of ethical and societal concerns, Governance implications of the use of artificial intelligence by organizations
- Overview:
  - Very high level endeavour
  - Requires domain-specific standardisation activities => language

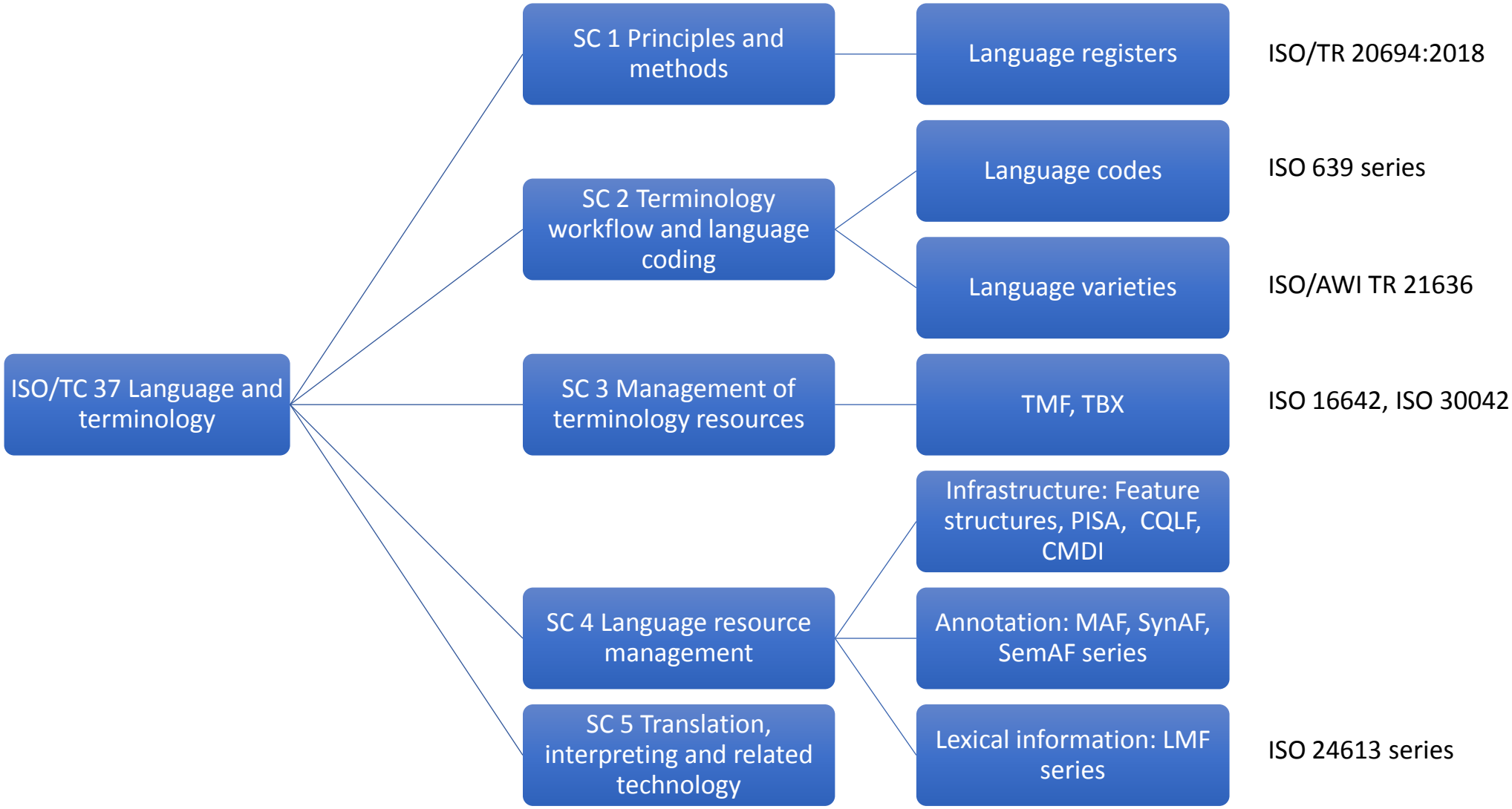
Language and standards – where?

- Various levels of specialisation
  - Horizontal vs. Vertical standards
- W3C - World Wide Web Consortium
  - Development of (horizontal) standards for the Internet
    - Ex. HTML, XML, Xpath, RDF, OWL, SKOS
    - The basic layer upon which language resources are being represented
- ISO - International Organization for Standardization
  - General purpose standardisation consortium
  - Organized in technical committees targeted at specific applications
    - ISO TC 37: Language and terminology (language codes, language resources, translation and interpretation)
- TEI - Text Encoding Initiative
  - Consortium dedicated to textual content
    - Drama, Transcription of spoken language, dictionaries, manuscripts, etc.
  - Based on XML – defines a vocabulary of around 600 elements

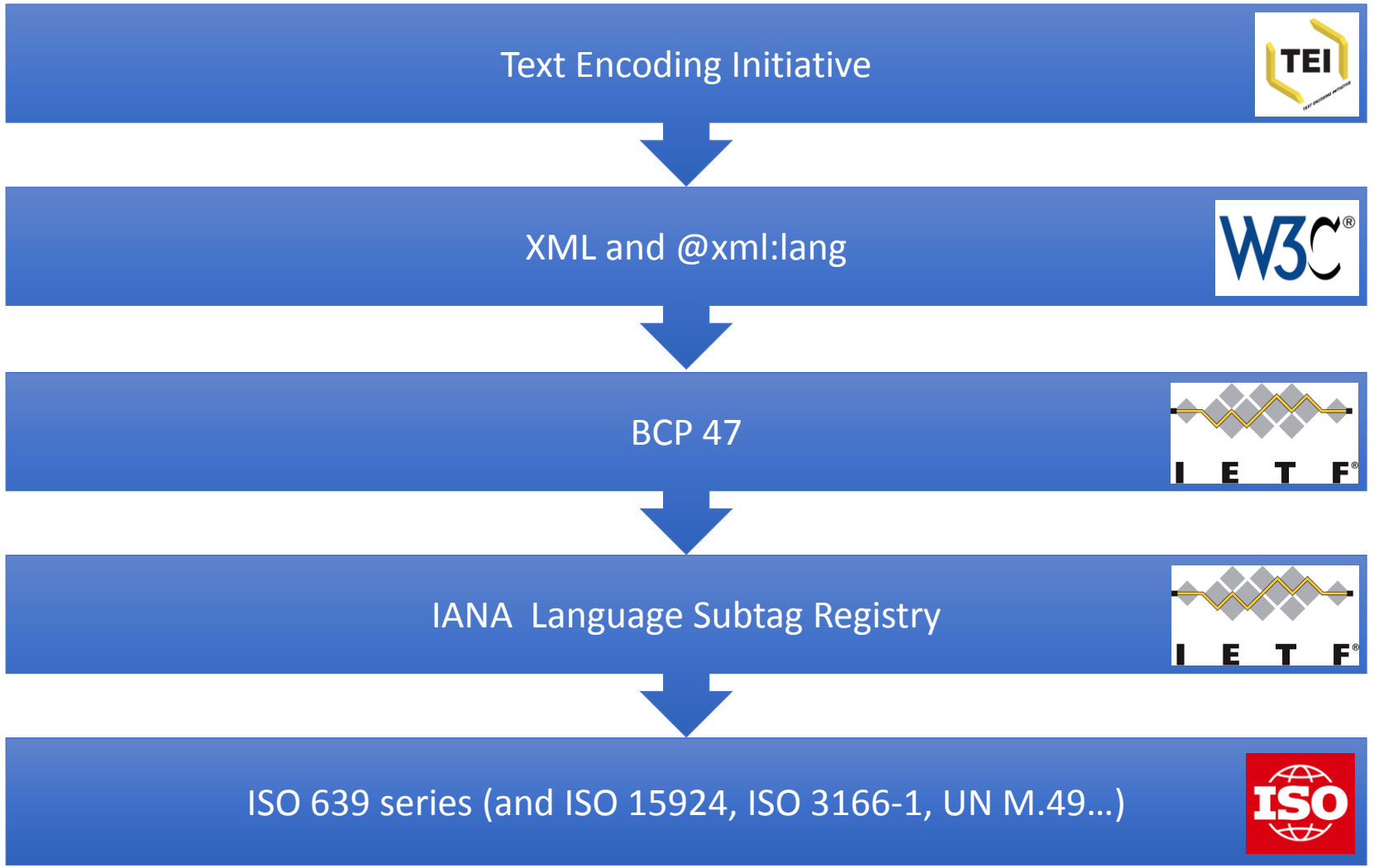
Language and standards – what and how?

- All levels of linguistic documentation, description, representation
  - Character encoding
  - Structure of documents
  - Linguistic annotation
  - Lexical resources
  - Metadata and bibliography
- The need for a variety of complementary standards
  - Standards piling

A selection of some ISO/TC 37 activities



Piling up standards




```
<entry xml:lang="fr-CA">
   <form>
      <orth>char</orth>
   </form>
```

```
xml:lang="fr-CA"
```

```
Fr-CA
```

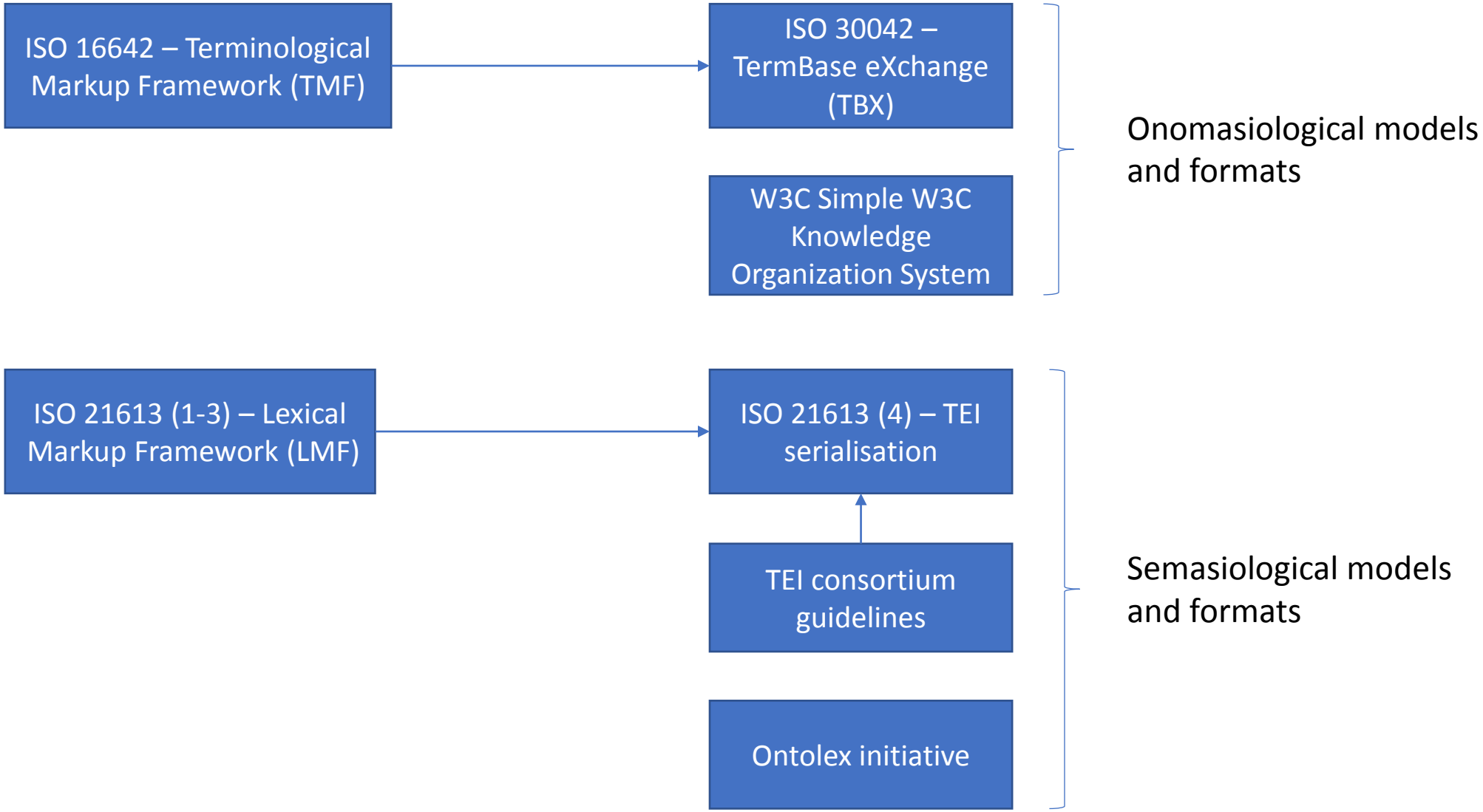
```
Type: language Subtag: fr
Description: French
```

```
fr, fra, fre
```

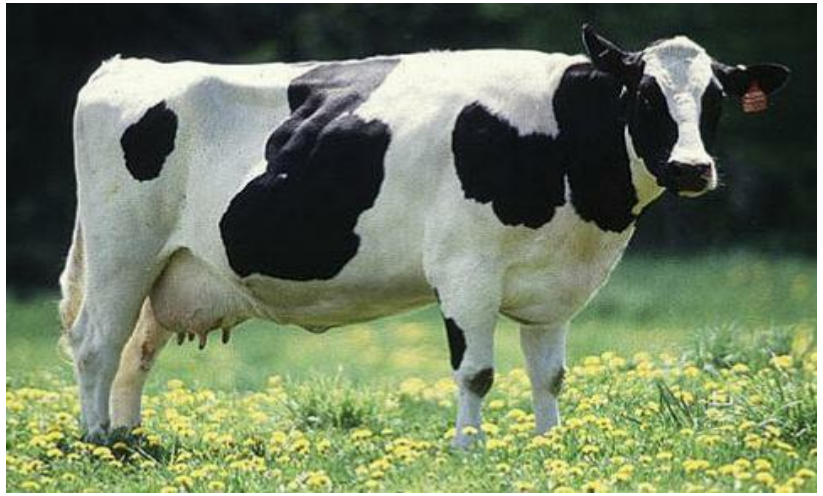
Dealing with variety: lexical resources

- A variety of contexts and forms
  - Legacy dictionaries, dialectological studies, NLP lexica
  - Onomasiological (concept to term) vs. semasiological (word to sense) structures
  - Lexical (machine processable) vs. Editorial (human readable) views
  - Full forms, etymology, corpus based examples
  - Word document, database, shoebox, XML…
- 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

Complementary of standardisation activities in the lexical domain



Back to machine learning: converting legacy dictionaries into TEI compliant structures



publishing

text-mining

Cow (structured data)

Hamburger (unstructured data)

"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)

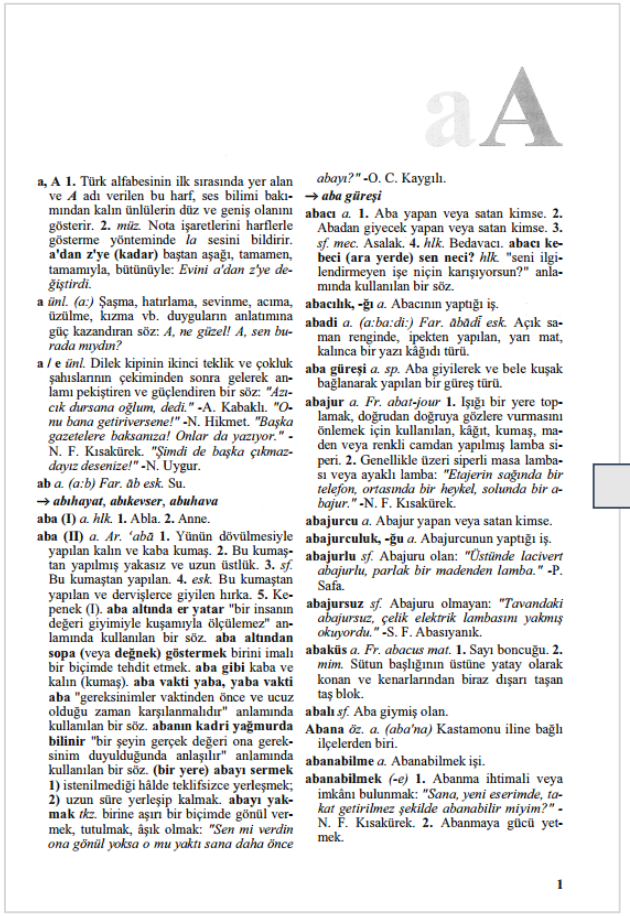
Inspired from: Duncan Hull

GROBID-Dictionaries

- Automatic extraction of TEI structures from digitised dictionaries (Khemakhem et al. 2017)
  - Input: PDF (or OCR/HTR formats such as 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

Cascading CRF models, an intuition

Cascading CRF models



a, A 1. Türk alfabesinin ilk sırasında yer alan ve *A* adı verilen bu harf, ses bilimi bakımından kalın ünlülerin düz ve geniş olanını gösterir. **2.** *müz.* Nota işaretlerini harflerle gösterme yönteminde *la* sesini bildirir. **a'dan z'ye (kadar)** baştan aşağı, tamamen, tamamıyla, bütünüyle: *Evini a'dan z'ye değiştirdi.*

a *ünl. (a:)* Şaşma, hatırlama, sevinme, acıma, üzülme, kızma vb. duyguların anlatımına güç kazandıran söz: *A, ne güzel! A, sen burada mıydın?*

a / e *ünl.* Dilek kipinin ikinci teklik ve çokluk şahıslarının çekiminden sonra gelerek anlamı pekiştiren ve güçlendiren bir söz: *"Azıcık dursana oğlum, dedi."* -A. Kabaklı. *"Onu bana getiriversene!"* -N. Hikmet. *"Başka gazetelere baksanıza! Onlar da yazıyor."* -N. F. Kısakürek. *"Şimdi de başka çıkmazdayız desenize!"* -N. Uygur.

ab *a. (a:b) Far. âb esk.* Su.

→ *abıhayat, abıkevser, abuhava*

aba (I) *a. hlk.* **1.** Abla. **2.** Anne.

aba (II) *a. Ar. 'abā* **1.** Yünün dövülmesiyle yapılan kalın ve kaba kumaş. **2.** Bu kumaştan yapılmış yakasız ve uzun üstlük. **3.** *sf.* Bu kumaştan yapılan. **4.** *esk.* Bu kumaştan yapılan ve dervişlerce giyilen hırka. **5.** Kepenek (I). **aba altında er yatar** "bir insanın değeri giyimiyle kuşamıyla ölçülemez" anlamında kullanılan bir söz. **aba altından sopa (veya değnek) göstermek** birini imalı bir biçimde tehdit etmek. **aba gibi** kaba ve kalın (kumaş). **aba vakti yaba, yaba vakti aba** "gereksinimler vaktinden önce ve ucuz olduğu zaman karşılanmalıdır" anlamında kullanılan bir söz. **abanın kadri yağmurda bilinir** "bir şeyin gerçek değeri ona gereksinim duyulduğunda anlaşılır" anlamında kullanılan bir söz. **(bir yere) abayı sermek 1)** istenilmediği hâlde teklifsizce yerleşmek; **2)** uzun süre yerleşip kalmak. **abayı yakmak** *tkz.* birine aşırı bir biçimde gönül vermek, tutulmak, âşık olmak: *"Sen mi verdin ona gönül yoksa o mu yaktı sana daha önce abayı?"* -O. C. Kaygılı.

→ *aba güreşi*

abacı *a.* **1.** Aba yapan veya satan kimse. **2.** Abadan giyecek yapan veya satan kimse. **3.** *sf. mec.* Asalak. **4.** *hlk.* Bedavacı. **abacı kebeci (ara yerde) sen neci?** *hlk.* "seni ilgilendirmeyen işe niçin karışıyorsun?" anlamında kullanılan bir söz.

abacılık, -ğı *a.* Abacının yaptığı iş.

abadi *a. (a:ba:di:) Far. âbâdî esk.* Açık saman renginde, ipekten yapılan, yarı mat, kalınca bir yazı kâğıdı türü.

aba güreşi *a. sp.* Aba giyilerek ve bele kuşak bağlanarak yapılan bir güreş türü.

abajur *a. Fr. abat-jour* **1.** Işığı bir yere toplamak, doğrudan doğruya gözlere vurmasını önlemek için kullanılan, kâğıt, kumaş, maden veya renkli camdan yapılmış lamba siperi. **2.** Genellikle üzeri siperli masa lambası veya ayaklı lamba: *"Etajerin sağında bir telefon, ortasında bir heykel, solunda bir abajur."* -N. F. Kısakürek.

abajurcu *a.* Abajur yapan veya satan kimse.

abajurculuk, -ğu *a.* Abajurcunun yaptığı iş.

abajurlu *sf.* Abajuru olan: *"Üstünde lacivert abajurlu, parlak bir madenden lamba."* -P. Safa.

abajursuz *sf.* Abajuru olmayan: *"Tavandaki abajursuz, çelik elektrik lambasını yakmış okuyordu."* -S. F. Abasıyanık.

abaküs *a. Fr. abacus mat.* **1.** Sayı boncuğu. **2.** *mim.* Sütun başlığının üstüne yatay olarak konan ve kenarlarından biraz dışarı taşan taş blok.

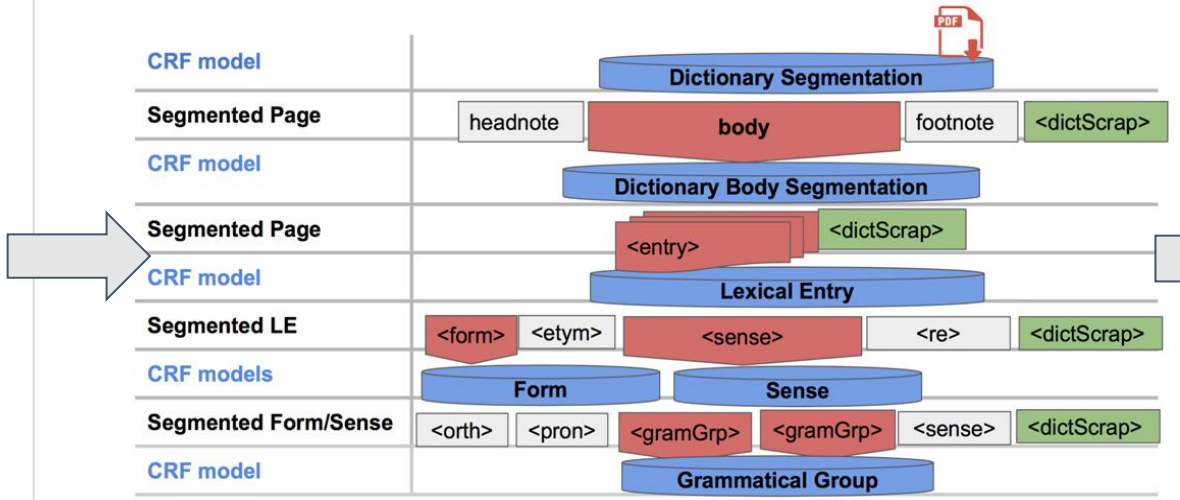
abalı *sf.* Aba giymiş olan.

Abana *öz. a. (aba'na)* Kastamonu iline bağlı ilçelerden biri.

abanabilme *a.* Abanabilmek işi.

abanabilmek *(-e)* **1.** Abanma ihtimali veya imkânı bulunmak: *"Sana, yeni eserimde, takat getirilmez şekilde abanabilir miyim?"* -N. F. Kısakürek. **2.** Abanmaya gücü yetmek.

1

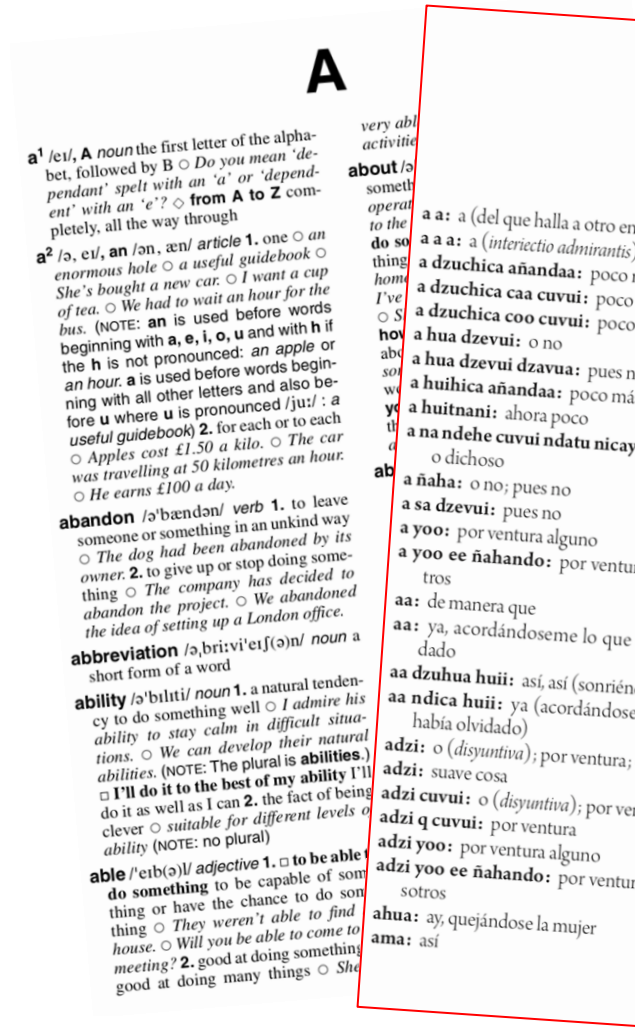


https://github.com/MedKhem/grobid-dictionaries

```
<entry>
   <form type="lemma" xml:id="center_form">
      <orth>center</orth>
      <pron>'sentɹ</pron>
      <gramGrp>
         <pos>noun</pos>
      </gramGrp>
      <usg type="geo">U.S</usg>
      <form type="variant">
         <orth>centre</orth>
         <usg type="geo">U.K</usg>
         <pron>'sentə</pron>
      </form>
   </form>
   <form type="inflected">
      <orth>centers</orth>
      <usg type="geo">U.S</usg>
      <gramGrp>
         <number>plural</number>
      </gramGrp>
   </form>
   <form type="inflected">
      <orth>centres</orth>
      <usg type="geo">U.K</usg>
      <gram type="number">plural</gram>
   </form>
   <sense>
      <def>the point around which a circle or sphere is described</def>
      <cit type="example">
         <quote>earth center"</quote>
      </cit>
   </sense>
   <sense>
      <gramGrp>
         <pos>verb</pos>
      </gramGrp>
      <def>place in the middle</def>
      <cit type="example">
         <quote>center the picture on the wall</quote>
      </cit>
   </sense>
   <re type="multiWordExpression">
      <form>
         <seg corresp="#dead_form" n="1">dead</seg>
         <seg corresp="#center_form" n="2">center</seg>
      </form>
   </re>
</entry>
```

Mohamed Khemakhem, Luca Foppiano, Laurent Romary. Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. *electronic lexicography, eLex 2017*, Sep 2017, Leiden, Netherlands. ⟨hal-01508868v2⟩

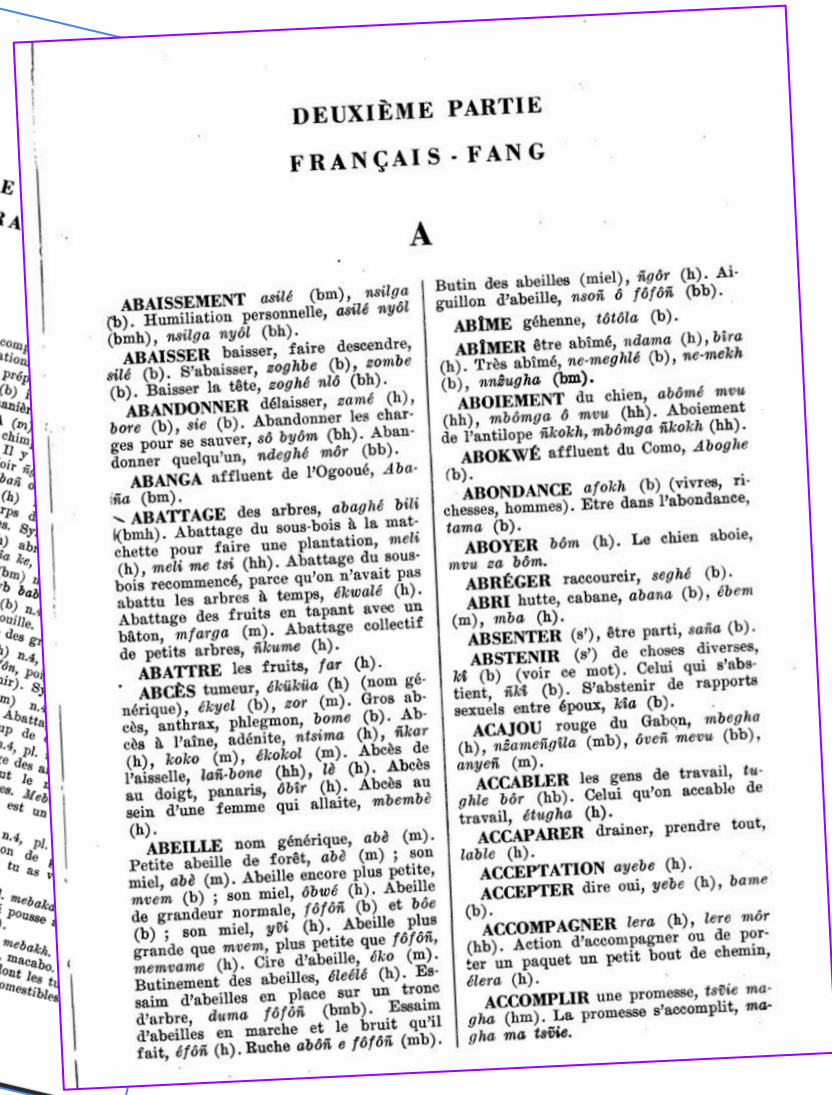
Experiments and results - 1



Easier English Basic Dictionary (2009)

Vocabulario en Lengua Mixteca (1593)

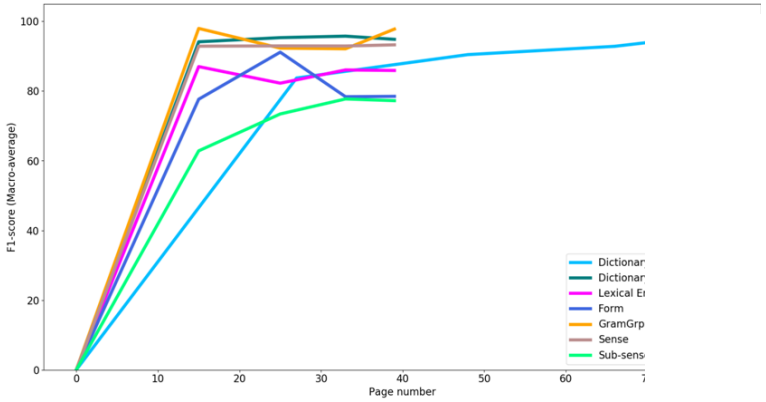
Dictionnaire de la langue Française (1873)



Dictionnaire fang-français (1964)

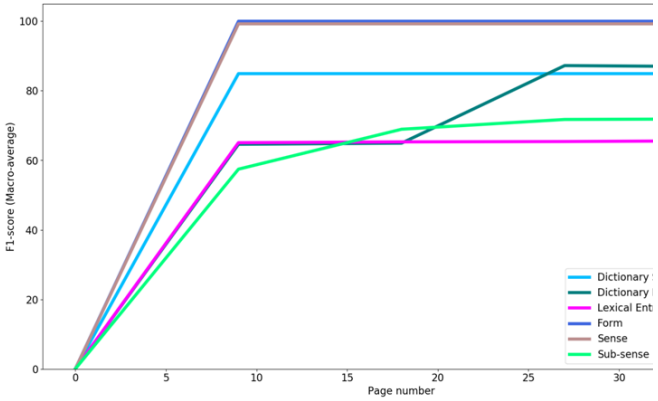
Experiments and results - 2

DLF

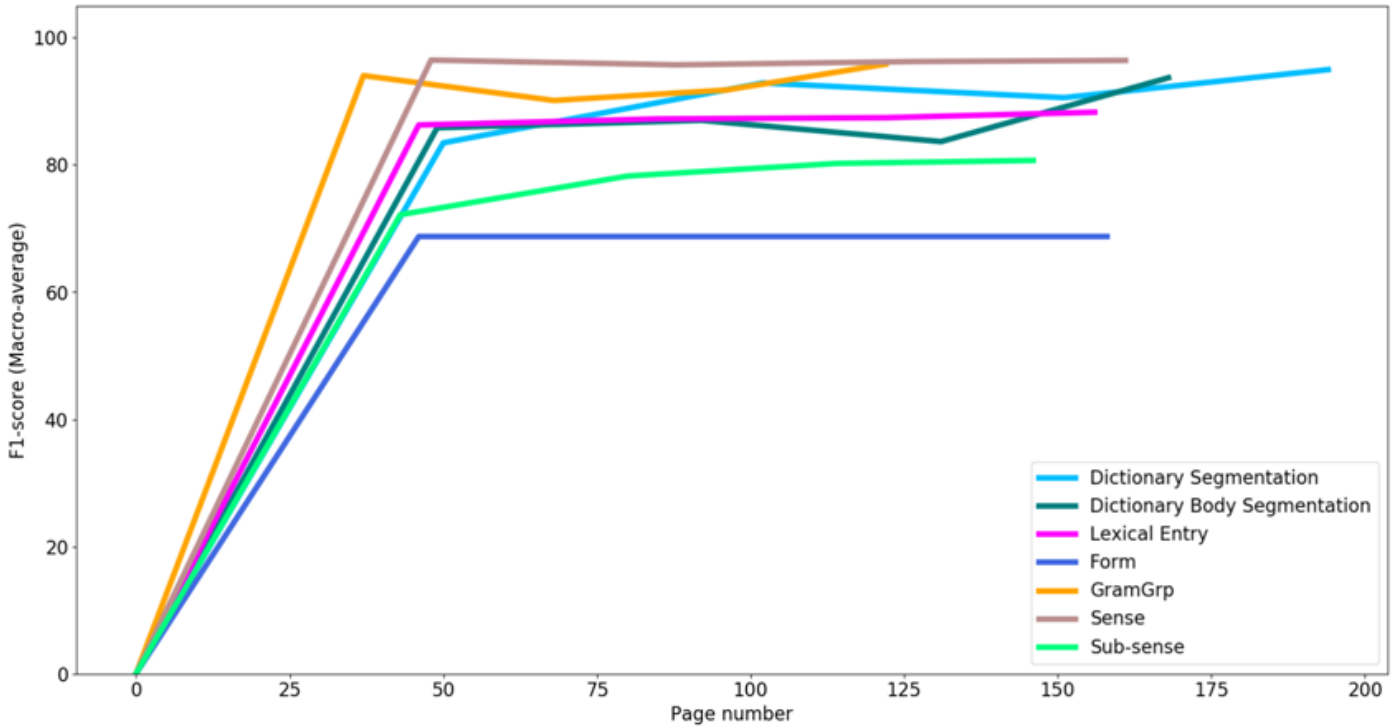


EEBD

MxSp



FrFang



FangFr

ALL

Further steps in language resources management

- Training
  - Basic digital literacy
    - Even for very simple issues related to online presence (authorship, affiliation)
  - Concrete knowledge of existing standards for language resources
    - Preventing the development of (any) new formats, unless strongly supported by evidence
- Active contributions
  - Participation to standardisation activities
  - Building up standards-based infrastructures
- General policies
  - A strong relation to open science principles (the I in FAIR, which leads to the R)
  - Hence a necessary stance for any research project

Merci pour votre attention