



**HAL**  
open science

# Multilingual content management and standards with a view on AI developments

Laurent Romary

► **To cite this version:**

Laurent Romary. Multilingual content management and standards with a view on AI developments. AI4EI - Conference Artificial Intelligence for European Integration, Oct 2020, Turin / Virtual, Italy. . hal-02961857

**HAL Id: hal-02961857**

**<https://inria.hal.science/hal-02961857v1>**

Submitted on 8 Oct 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Multilingual content management and standards with a view on AI developments

Laurent Romary

Directeur de Recherche, Inria, team ALMAAnaCH

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



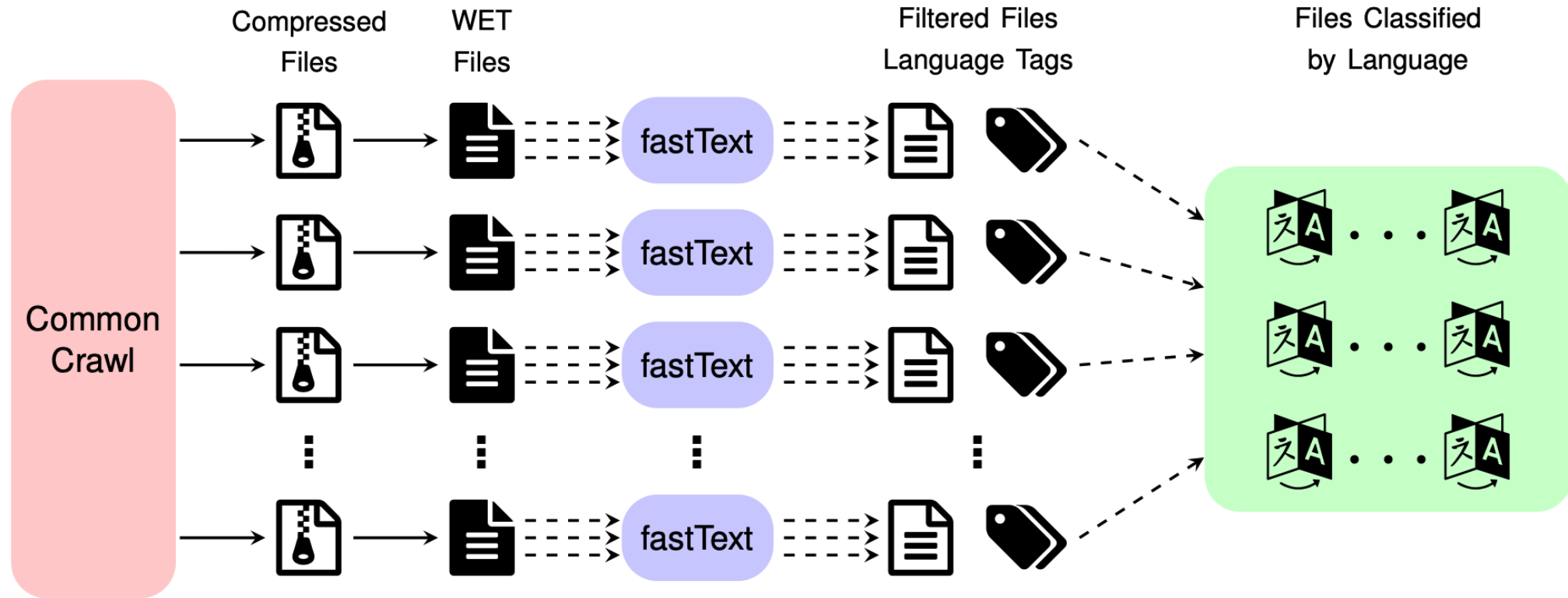
Language model



CamemBERT

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*.

# The infrastructure to create OSCAR - goclassy



# 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
<b>Total OSCAR</b>	<b>6.3T</b>	<b>3.2T</b>	<b>844,315,434,723</b>	<b>425,651,344,234</b>

# 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

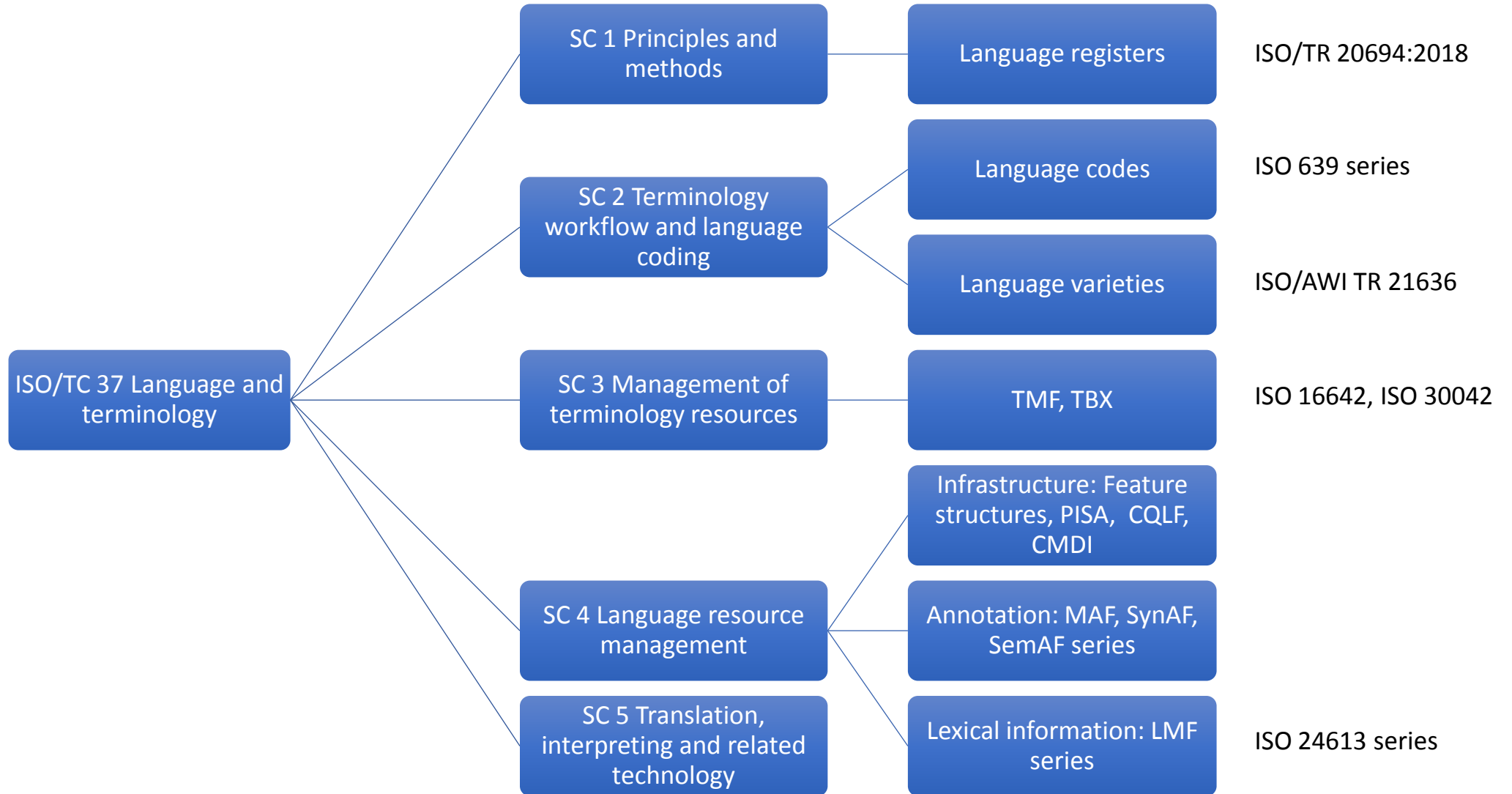
# Language and standards – where?

- 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

Text Encoding Initiative



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

XML and @xml:lang



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

BCP 47



```
Fr-CA
```

IANA Language Subtag Registry



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

ISO 639 series (and ISO 15924, ISO 3166-1, UN M.49...)

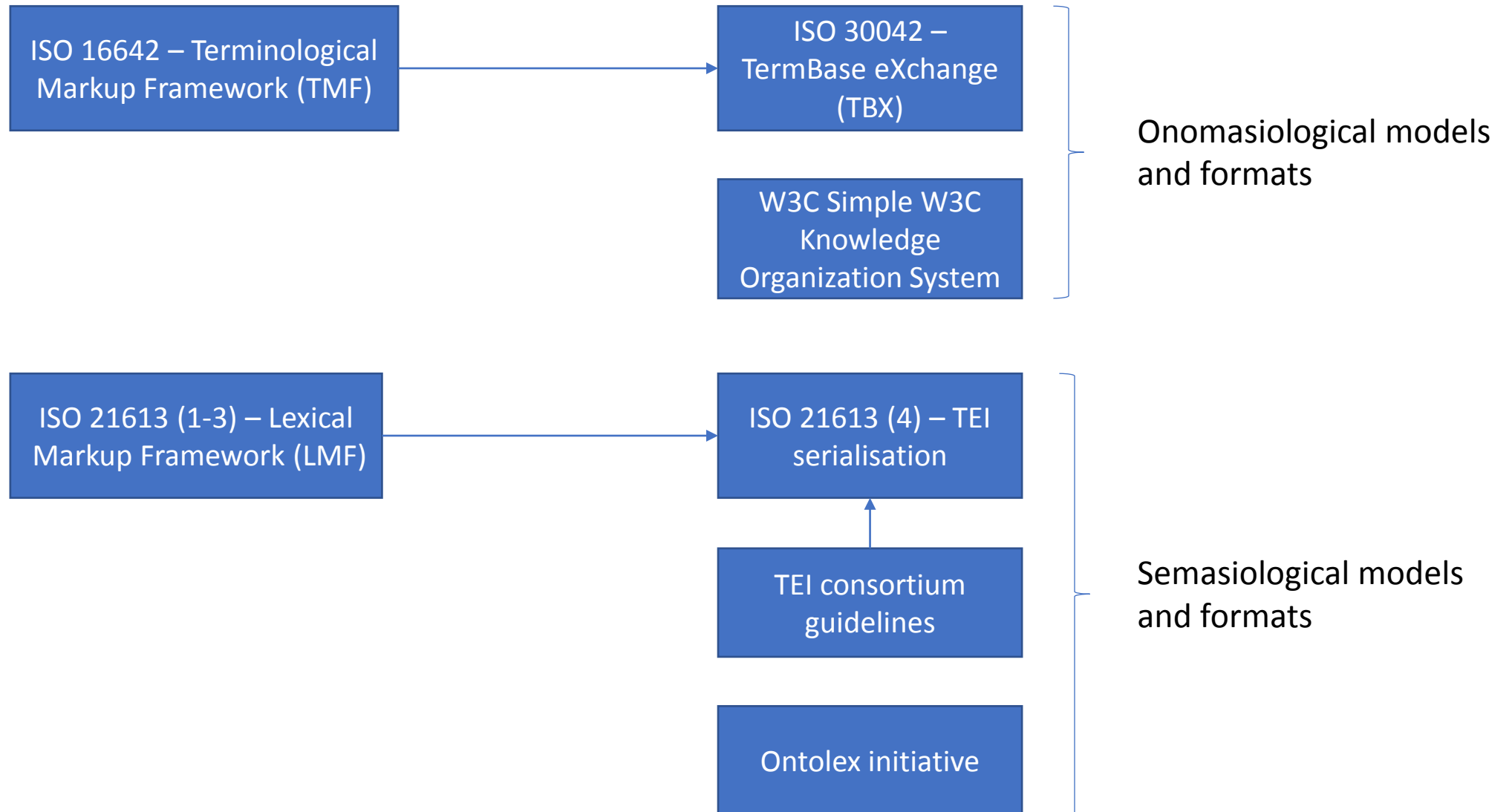


```
fr, fra, fre
```

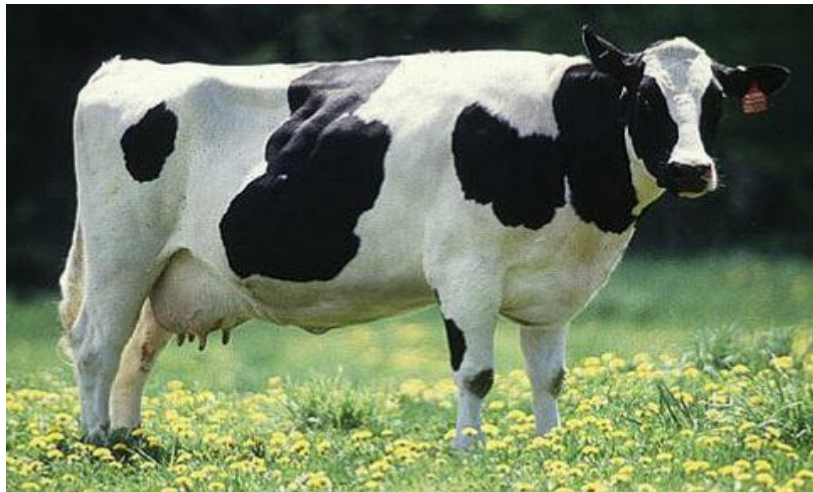
# Dealing with variety: lexical resources

- 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

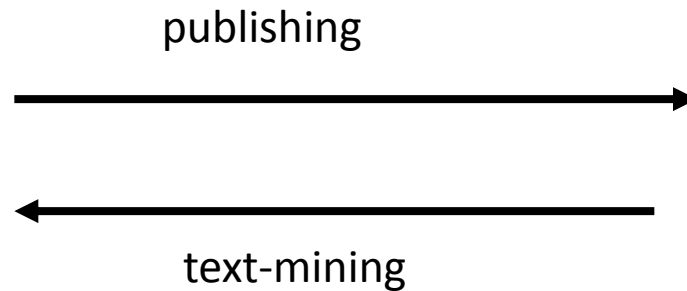
# Complementary of standardisation activities in the lexical domain



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



**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>)



# 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

Eugène IV à Ferrare en 1438-1439, puis à Florence de 1439 à 1442), de Latran (1512-1517), de Trente (1545-1563) [où fut décidée la réforme générale de l'Église catholique en face de la Réforme protestante], de Vatican I (1870) [où fut défini le dogme de l'infaillibilité pontificale], de Vatican II (1962-1965) [où fut définie l'attitude de l'Église romaine à l'égard du monde moderne].  
**conciliable** adj. Qui peut se concilier avec une autre chose.  
**conciliabule** [kɔ̃siljabyl] n. m. (lat. conciliabulum).

**concordant, e** adj. Qui s'accorde : *témoignages concordants*.  
**concordat** [kɔ̃kɔrdɑ] n. m. (lat. concordatum). Traité entre le pape et un gouvernement sur les affaires religieuses. || *Dr.* Accord entre le commerçant qui, ayant déposé son bilan, a été admis par le tribunal de commerce au règlement judiciaire et ses créanciers. — Les plus anciens concordats sont le concordat de Worms (1122), entre Calixte II et Henri V; le concordat de 1516, entre Léon X et François I<sup>er</sup>. Le concordat entre

**concupiscence** n. f. (du lat. concupiscere, désirer). Penchant à jouir des biens terrestres, particulièrement des plaisirs sensuels.  
**concupiscent** [kɔ̃kypisɑ̃] n. m. (lat. concupiscens). Attaché aux plaisirs sensuels.  
**concurrer** [kɔ̃kyrɑ̃] v. t. (lat. concurrere). Par concurrence. Par un concours mutuel, de concert : *agir concurremment avec quelqu'un*.  
**concurrance** n. f. Rivalité entre plusieurs personnes qui visent un même but : *entrer*

**condenser** [kɔ̃dɑ̃sɛ] v. t. (lat. condensare, rendre épais). **Rendre plus dense, réduire à un moindre volume. || Liquéfier un gaz par refroidissement ou compression : le froid condense la vapeur d'eau. || Fig. Exprimer d'une manière concise, en peu de mots :**

**conclaviste** n. m. Personne qui s'enferme au conclave avec un cardinal, pour le servir.  
**concluant, e** adj. Qui prouve bien ce qu'on a avancé : *argument concluant*.  
**conclure** [kɔ̃klyʁ] v. t. (lat. concludere) [con]. 62. Achever, terminer : *conclure une affaire*. || Tirer une conséquence : *conclure une chose d'une autre*. || — V. l. Donner son avis, ses conclusions; se prononcer : *on vous demande de conclure*. || Être proham : *les témoignages concluent contre lui*.  
**conclusion** n. f. (lat. conclusio). Arrangement définitif : *conclusion d'un traité*. || Fin, résultat final : *la conclusion d'un discours*. || Conséquence d'un argument : *la conclusion d'un syllogisme ne doit pas dépasser les prémisses*. || — Pl. Prétentions respectives de chacune des parties dans un procès. || Écrit exposant ces prétentions. || Réquisition du ministère public. || — En conclusion loc. adv. En conséquence, pour conclure.  
**concocter** v. t. Fam. Elaborer avec soin : *concocter une lettre de réclamation*.  
**concombre** [kɔ̃kɔbr] n. m. (anc. provenç. cocombre). Plante potagère de la famille des cucurbitacées, cultivée pour ses fruits allongés que l'on consomme comme légume ou en salade. || Ce fruit.  
**concomitamment** adv. De façon concomitante.  
**concomitance** [kɔ̃kɔmitɑ̃] n. f. Coexistence, simultanéité de deux ou de plusieurs faits.  
**concomitant, e** adj. (lat. concomitans). Qui accompagne, qui se produit en même temps : *des faits concomitants*. • *Variations concomitantes*, variations simultanées et proportionnelles de certains phénomènes.  
**concordance** n. f. Conformité, accord : *concordance de témoignages*. || Géol. Disposition parallèle des couches sédimentaires. • *Concordance de phases* (Phys.), état de plusieurs vibrations sinusoïdales de même nature et de même période, dont la différence de phases est nulle. || *Concordance des temps*, règles de syntaxe d'après lesquelles le temps du verbe d'une subordonnée varie selon celui du verbe de la principale.

**concours**. || Lutte sportive : *concours hippique*. • *Concours général*, concours annuel entre les premiers élèves des classes supérieures des lycées, collèges et écoles normales.  
**concret** [kɔ̃kʁɛ] n. m. (lat. concretus). Epais, condensé : *huile concrète* (vieille). || Qui exprime quelque chose de réel, de positif : *obtenir des avantages concrets*. || Qui a le sens des réalités précises : *esprit concret*. || Gramm. Se dit d'un terme qui désigne un être ou un objet pouvant être perçu par les sens. • *Musique concrète*, technique de composition qui utilise les bruits produits par divers objets sonores enregistrés sur bande magnétique et susceptibles de transformation.  
**concret** n. m. Qualité de ce qui est concret.  
**concrètement** adv. De façon concrète.  
**concréter** v. t. (conç. 3). Rendre concret, solide.  
**concrétion** [kɔ̃kʁɛsjɔ̃] n. f. (de concret). Action de s'épaissir : *la concrétion de l'huile, du sang*. || Réunion de parties en un corps solide : *concrétion saline*. || Agrégation solide dans les tissus vivants : *concrétions biliaires*.  
**concrétiser** v. t. Rendre concret ce qui est abstrait : *concrétiser une idée, un avantage*.  
**concubin, e** adj. (lat. concubina). Relatif au concubinage. || — N. Personne qui vit en concubinage.  
**concubinage** [kɔ̃kybinɑ̃] n. m. État d'un homme et d'une femme qui vivent ensemble sans être mariés. (On dit aussi UNION LIBRE.)

**condensation**.  
**condamnabile** adj. Qui mérite d'être condamné : acte *condamnabile*.  
**condamnation** [kɔ̃dɑ̃nasjɔ̃] n. f. (lat. condemnatio). Décision d'un tribunal important à l'un des plaideurs de s'incliner au moins partiellement devant les prétentions de son adversaire. || Décision d'une juridiction prononçant une peine contre l'auteur d'un crime, d'un délit ou d'une contravention. (En cour d'assises, le jury juge la culpabilité de l'accusé, et la cour prononce la condamnation.) || La peine infligée : *une condamnation à la réclusion criminelle*. || Blâme, désapprobation : *la condamnation des abus*.  
**condamnatore** adj. Qui porte condamnation.  
**condamné, e** n. Personne qui a subi une condamnation. || — Adj. Qui ne peut échapper à un sort prévu : *malade condamné*.  
**condamner** [kɔ̃dɑ̃nɛ] v. t. (lat. condemnare). Prononcer un jugement contre un plaideur ou un inculpé : *condamner un criminel*. || Astreindre, réduire à : *condamner au silence, à l'immobilité*. || Désapprouver, blâmer : *condamner une opinion, un usage*. || Interdire : *la loi condamne la bigamie*. || Déclarer perdu, incurable : *les médecins l'ont condamné*. || Barre, murer : *condamner une porte*.  
**condensable** adj. Qui peut être condensé, réduit à un moindre volume.  
**condensateur** n. m. Phys. Appareil servant à emmagasiner une charge électrique : *la bouteille de Leyde est un condensateur électrique*. || Lentille servant à éclairer un objet dont on veut former une image.  
**condensation** n. f. Action de condenser ou effet qui en résulte. || Liquéfaction d'un gaz. || Soudure de plusieurs molécules chimiques, avec élimination d'eau.  
**condensé** n. m. Résumé d'une œuvre littéraire.  
**condenser** [kɔ̃dɑ̃sɛ] v. t. (lat. condensare, rendre épais). Rendre plus dense, réduire à un moindre volume. || Liquéfier un gaz par refroidissement ou compression : *le froid condense la vapeur d'eau*. || Fig. Exprimer d'une manière concise, en peu de mots :



# Cascading CRF models

aA

**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. *miz*. Nota işaretlerini harflerle gösterme yönteminde *la* sesini bildirir. *a'dan z'ye (kadar)* baştan aşağı, tamamen, tamamıyla, bütünüyle: *Evinin a'dan z'ye de-ğiştirildi.*

**a ünl.** (*a*) Şaşma, hatırlama, sevinme, acıma, üzülmeye, 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 gelecek anlamı pekiştiren ve güçlendiren bir söz: *"Azıcık dursana oğlum, dedi." -A. Kabaklı. "Başka gazetelere baksanıza! Onlar da yazıyor." -N. F. Kısakürek. "Şimdi de başka çıkmazdayız deseniz!" -N. Uygur.*

**ab a.** (*a:b*) *Far. ab* esk. Su.  
→ *abıhayat, abıkevser, abıhava*

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

**aba (II) a. Ar.** 'abâ 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 derişlerce giyilen hırka. 5. Ke-penek (I). *aba altında er yatar* "bir insanın değeri giyimiyle kumaşıyla 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 karşılmalıdır" 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 yapmak* *öz.* birine aşırı bir biçimde gönlül vermek, tutulmak, âşık olmak: *"Sen mi verdin ona gönlül yoksa o mu yaptı sana daha önce*

*abaya?*" -O. C. Kaygılı.  
→ *aba gürüşü*

**abacı a.** 1. *Abacı* 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. abâdi* esk. Açık saman renginde, ipekten yapılan, yarı mat, kalınca bir yazı kâğıdı türü.

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

**abajur a. Fr. abat-jour** 1. İşığı bir yere toplamak, doğrudan doğruya gözle vurmaması önlemek için kullanılan, kâğıt, kumaş, maden veya renkli camdan yapılmış lamba siperi. 2. Genellikle üzeri siperli masa lambası veya ayaklı lamba: *"Etuajın 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.** *Abajurlu* olan: *"Üstünde lacivert abajurlu, parlak bir madenden lamba."* -P. Safa.

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

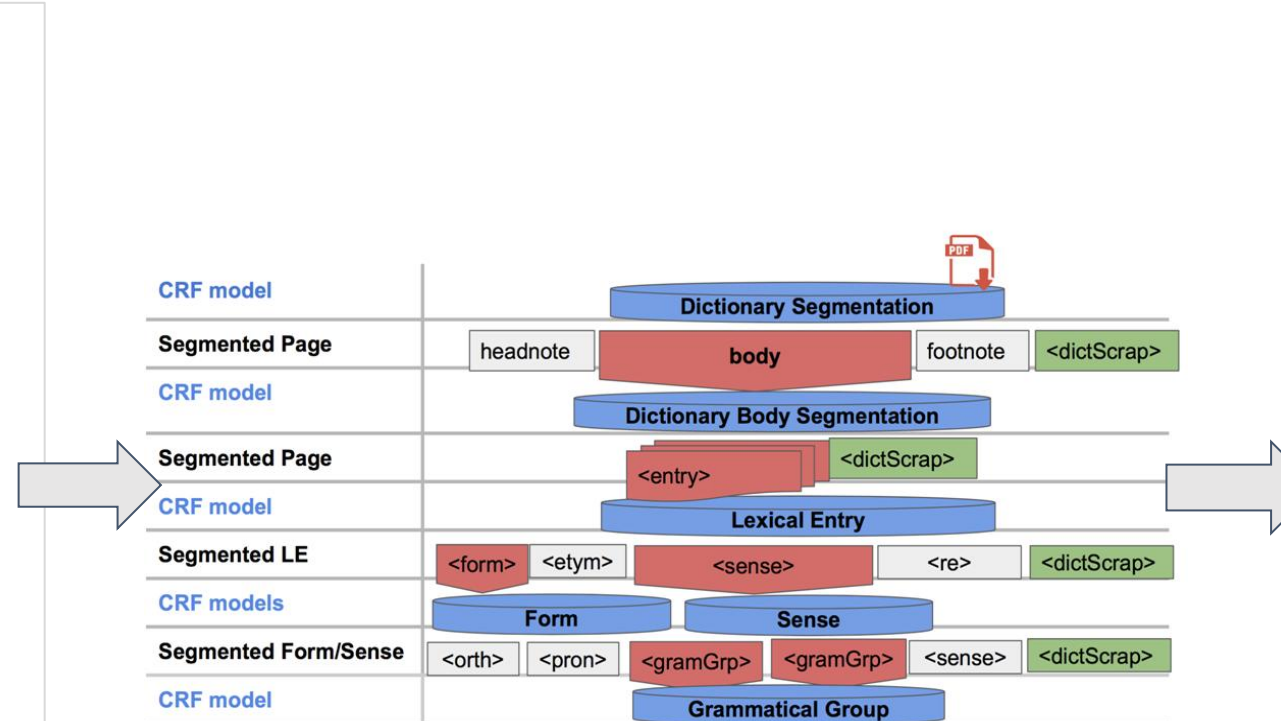
**abaküs a. Fr. abacus** *mat.* 1. Sayı boncuğu. 2. *mim.* Sütün başlığının üstüne yatay olarak konan ve kenarlarından biraz dışarı taşan taş blok.

**abalı sf.** *Aba* giymiş olan.

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

**abanabilme a.** *Abanabilmek işi.*

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



```

<entry>
  <form type="lemma" xml:id="center_form">
    <orth>center</orth>
    <pron>'sent</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>'sent</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>

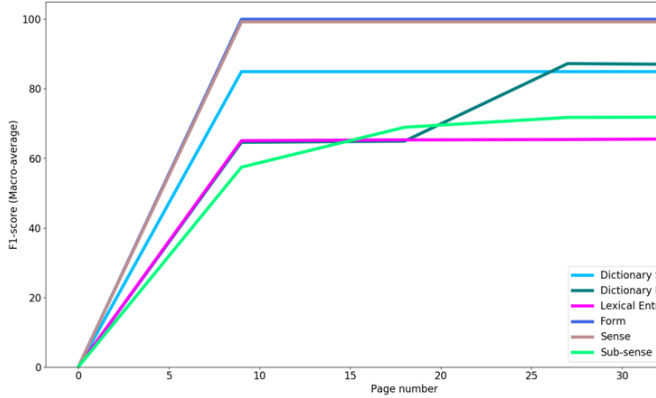
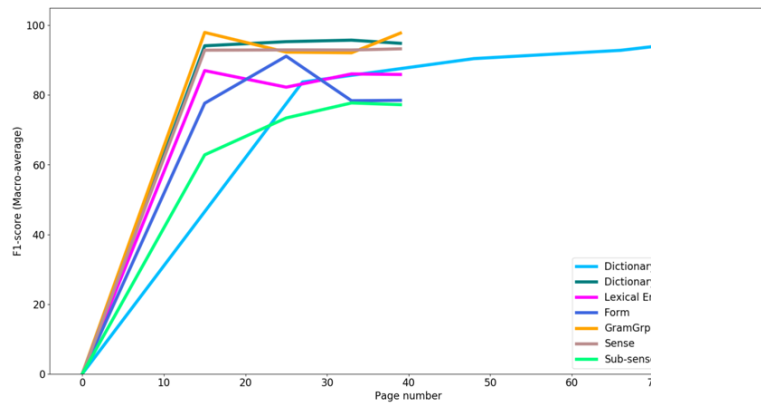
```

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



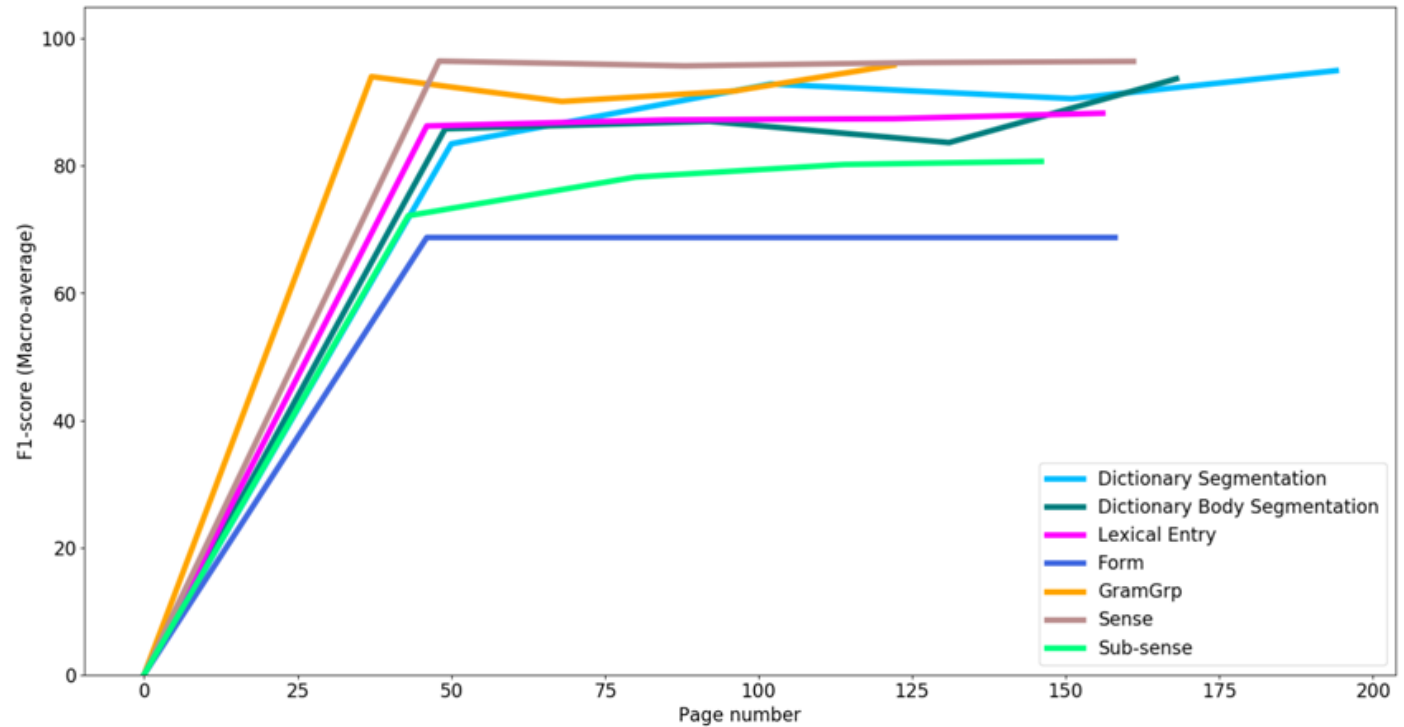
# Experiments and results - 2

DLF



FrFang

EEBD



FangFr

MxSp

**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