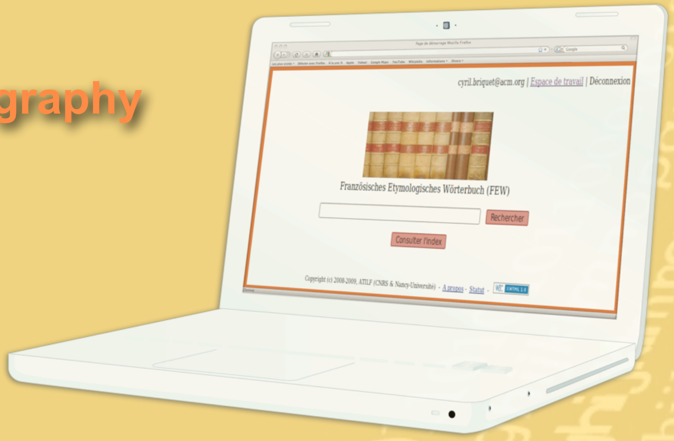
Retroconversion Of A Complex Etymological Dictionary

European Master in Lexicography

2009-2010



cyril.briquet@acm.org | Espace de travail | Déconnexion

Französisches Etymologisches Wörterbuch (FEW)

Copyright (c) 2008-2009, ATILF (CNRS & Nancy-Université) - A propos - Statut -

Pascale Renders (1) (2), Cyril Briquet (1)
(1) ATILF (CNRS & Nancy-Université), (2) Université de Liège
pascale.renders@atilf.fr, cyril.briquet@acm.org

http://www.atilf.fr/few

Outline

A. Presentation of the Project

B. The Retroconversion System

C. Beyond Retroconversion

Outline

A. Presentation of the Project

1. The FEW
2. Retroconverting the FEW
3. Exploitation

A.1. The FEW

Französisches Etymologisches Wörterbuch

- Walther von Wartburg
- 25 volumes published from 1922 to 2002, in German and French
- *Thesaurus galloromanicus*
- French, Franco-provençal, Occitan, Gascon in all their diatopic variations, from IX^th^ to XX^th^ century
- Etymology-history (genetic perspective)

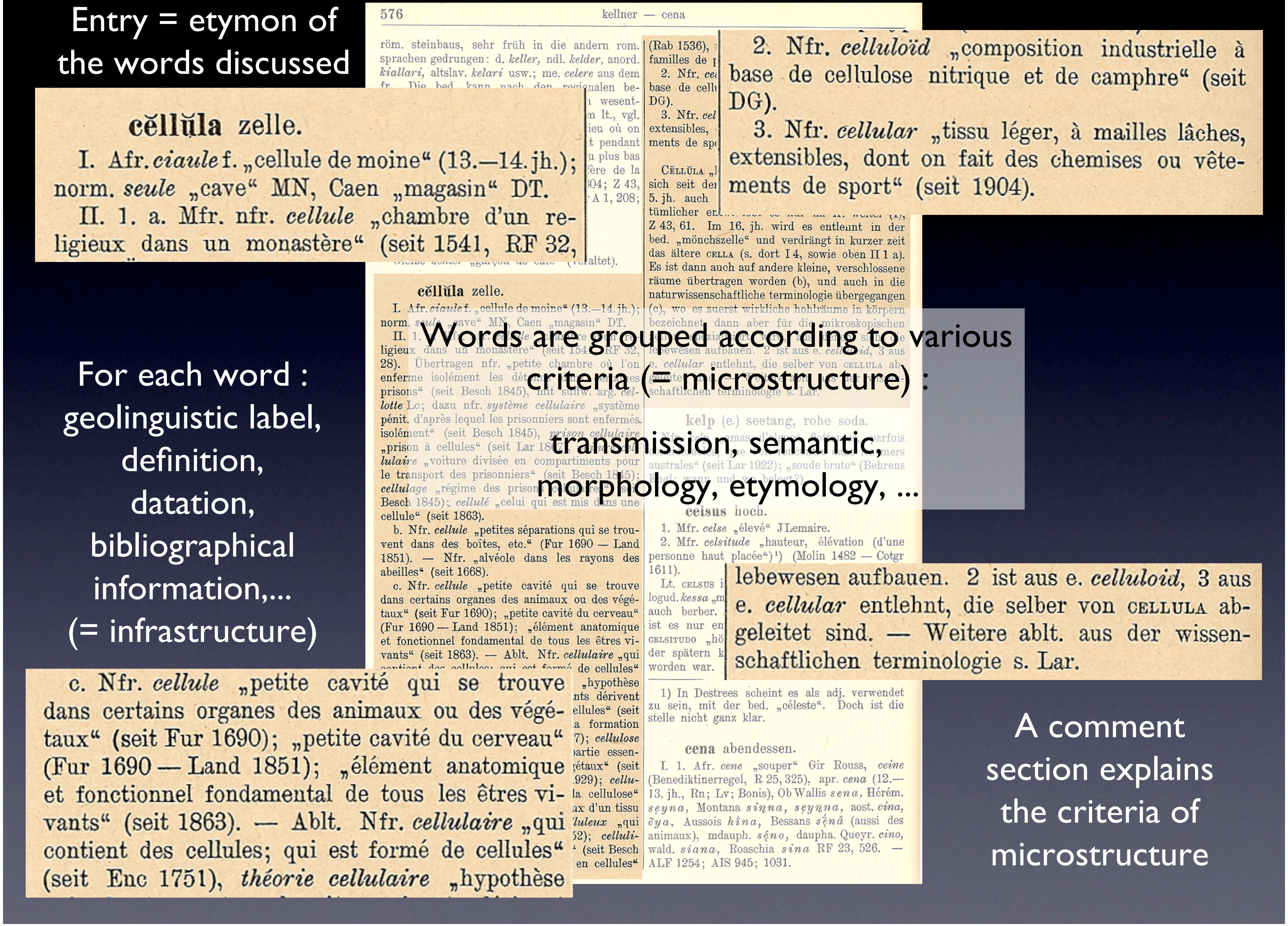
Entry = etymon of the words discussed

For each word : geolinguistic label, definition, datation, bibliographical information,... (= infrastructure)

Words are grouped according to various criteria (= microstructure) :

transmission, semantic, morphology, etymology, ...

A comment section explains the criteria of microstructure



576 kellner — cena

röm. steinbaus, sehr früh in die andern rom. sprachen gedrungen: d. *keller*, ndl. *kelder*, anord. *kiallari*, altslav. *kelari* usw.; me. *celere* aus dem fr. Die bed. kann nach den regionalen be- … n wesent- … m lt., vgl. … ieu où on … t pendant … u plus bas … fère de la … 04; Z 43, … A 1, 208; … (veraltet).

cĕllŭla zelle.

I. Afr. *ciaule* f. „cellule de moine“ (13.—14. jh.); norm. *seule* „cave“ MN, Caen „magasin“ DT.

II. 1. a. Mfr. nfr. *cellule* „chambre d'un religieux dans un monastère“ (seit 1541, RF 32, 28). Übertragen nfr. „petite chambre où l'on enferme isolément les détenus dans certaines prisons“ (seit Besch 1845), mit suffw. arg. *cellotte* Lc; dazu nfr. *système cellulaire* „système pénit. d'après lequel les prisonniers sont enfermés isolément“ (seit Besch 1845), *prison cellulaire* „prison à cellules“ (seit Lar 18[illegible]), *voiture cellulaire* „voiture divisée en compartiments pour le transport des prisonniers“ (seit Besch 1845); *cellulage* „régime des prisons cellulaires“ (seit Besch 1845); *cellulé* „celui qui est mis dans une cellule“ (seit 1863).

b. Nfr. *cellule* „petites séparations qui se trouvent dans des boîtes, etc.“ (Fur 1690 — Land 1851). — Nfr. „alvéole dans les rayons des abeilles“ (seit 1668).

c. Nfr. *cellule* „petite cavité qui se trouve dans certains organes des animaux ou des végétaux“ (seit Fur 1690); „petite cavité du cerveau“ (Fur 1690 — Land 1851); „élément anatomique et fonctionnel fondamental de tous les êtres vivants“ (seit 1863). — Ablt. Nfr. *cellulaire* „qui contient des cellules; qui est formé de cellules“ (seit Enc 1751), *théorie cellulaire* „hypothèse … nts dérivent … ellules“ (seit … a formation … 7); *cellulose* … partie essen- … gétaux“ (seit … 929); *cellu-* … la cellulose“ … ax d'un tissu … *lluleux* „qui … 52); *celluli-* … “ (seit Besch … en cellules“

(Rab 1536), … familles de p…

2. Nfr. *celluloïd* „composition industrielle à base de cellulose nitrique et de camphre“ (seit DG).

3. Nfr. *cellular* „tissu léger, à mailles lâches, extensibles, dont on fait des chemises ou vêtements de sport“ (seit 1904).

Cĕllŭla „… sich seit der … 5. jh. auch … tümlicher en… Z 43, 61. Im 16. jh. wird es entleunt in der bed. „mönchszelle“ und verdrängt in kurzer zeit das ältere cella (s. dort I 4, sowie oben II 1 a). Es ist dann auch auf andere kleine, verschlossene räume übertragen worden (b), und auch in die naturwissenschaftliche terminologie übergegangen (c), wo es zuerst wirkliche hohlräume in körpern bezeichnet, dann aber für die mikroskopischen … spezialisiert wurde, aus denen sich die lebewesen aufbauen. 2 ist aus e. *celluloid*, 3 aus e. *cellular* entlehnt, die selber von cellula abgeleitet sind. — Weitere ablt. aus der wissenschaftlichen terminologie s. Lar.

kelp (e.) seetang, rohe soda.

Nfr. *kelp* „amas d'algues flottantes, parfois considérables, que l'on rencontre dans les mers australes“ (seit Lar 1922); „soude brute“ (Behrens Engl.; wann und wo belegt?).

celsus hoch.

1. Mfr. *celse* „élevé“ JLemaire.

2. Mfr. *celsitude* „hauteur, élévation (d'une personne haut placée“)[1] (Molin 1482 — Cotgr 1611).

Lt. celsus i… logud. *kessa* „m… auch berber. … ist es nur en… celsitudo „hö… der spätern k… worden war.

[1] In Destrees scheint es als adj. verwendet zu sein, mit der bed. „céleste“. Doch ist die stelle nicht ganz klar.

cena abendessen.

I. 1. Afr. *cene* „souper“ Gir Rouss, *ceine* (Benediktinerregel, R 25, 325), apr. *cena* (12.—13. jh., Rn; Lv; Bonis), Ob Wallis *sena*, Hérém. *sẹyna*, Montana *siŋna*, *sẹyŋna*, aost. *cina*, *ẽya*, Aussois *hĩna*, Bessans *sę̃nå* (aussi des animaux), mdauph. *sẹ́no*, daupha. Queyr. *cino*, wald. *síana*, Roaschia *sina* RF 23, 526. — ALF 1254; AIS 945; 1031.

cĕllŭla zelle.

I. Afr. *ciaule* f. „cellule de moine“ (13.—14. jh.); norm. *seule* „cave“ MN, Caen „magasin“ DT.

II. 1. a. Mfr. nfr. *cellule* „chambre d'un religieux dans un monastère“ (seit 1541, RF 32,

2. Nfr. *celluloïd* „composition industrielle à base de cellulose nitrique et de camphre“ (seit DG).

3. Nfr. *cellular* „tissu léger, à mailles lâches, extensibles, dont on fait des chemises ou vêtements de sport“ (seit 1904).

lebewesen aufbauen. 2 ist aus e. *celluloid*, 3 aus e. *cellular* entlehnt, die selber von cellula abgeleitet sind. — Weitere ablt. aus der wissenschaftlichen terminologie s. Lar.

c. Nfr. *cellule* „petite cavité qui se trouve dans certains organes des animaux ou des végétaux“ (seit Fur 1690); „petite cavité du cerveau“ (Fur 1690 — Land 1851); „élément anatomique et fonctionnel fondamental de tous les êtres vivants“ (seit 1863). — Ablt. Nfr. *cellulaire* „qui contient des cellules; qui est formé de cellules“ (seit Enc 1751), *théorie cellulaire* „hypothèse

Structural Complexity

Reference book in French and Romance Linguistics (along with LEI for Italian dialects), but...

not easy to read, because of

- its complex structure
- the large number of informational fields
- the implicitness of its content, both syntactic (abbreviations) and semantic

not easy to search :
searching for specific kinds of words in the whole dictionary is not possible !

Challenges

These issues (readability, transversal search) could certainly be adressed

1. if the FEW were computerized
2. and if its contents were semantically searchable.

An exciting question is :

starting from the printed version of the dictionary, how can the complex dictionary structure be extracted into a searchable database ?

A.2. Retroconverting The FEW

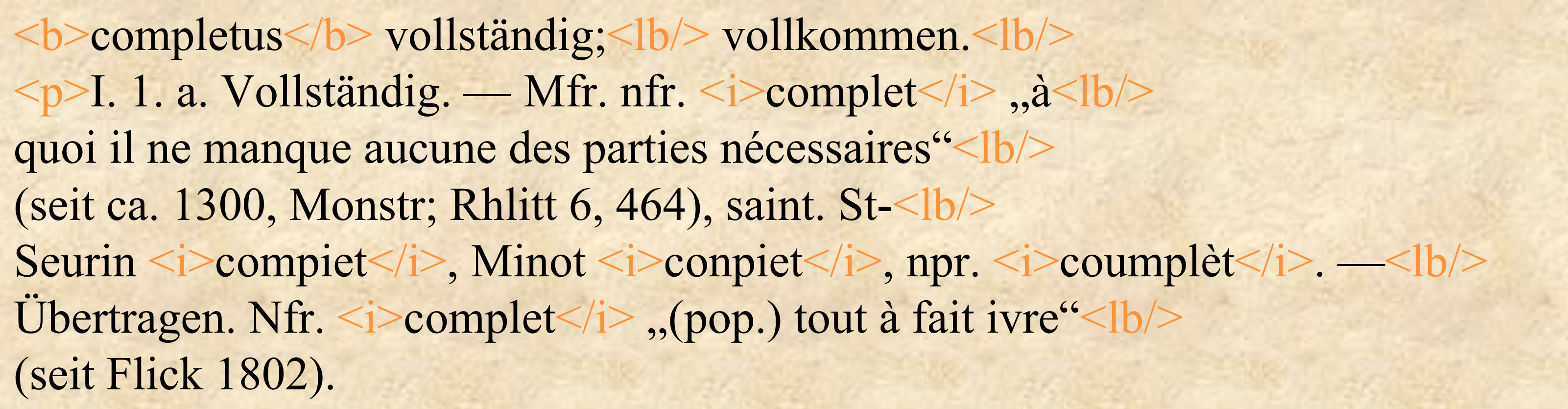
What is retroconversion ?

Computerizing a paper dictionary consists in turning it into a dictionary digitized to a certain extent :

- image files : scan pages to provide raw visual contents
- plain text files : ocerize (OCR) to provide raw textual contents
- domain-specific XML files : perform analysis to provide semantically-structured contents

To be tractable, the retroconversion process should be as automated as possible.

FEW Retroconversion Input

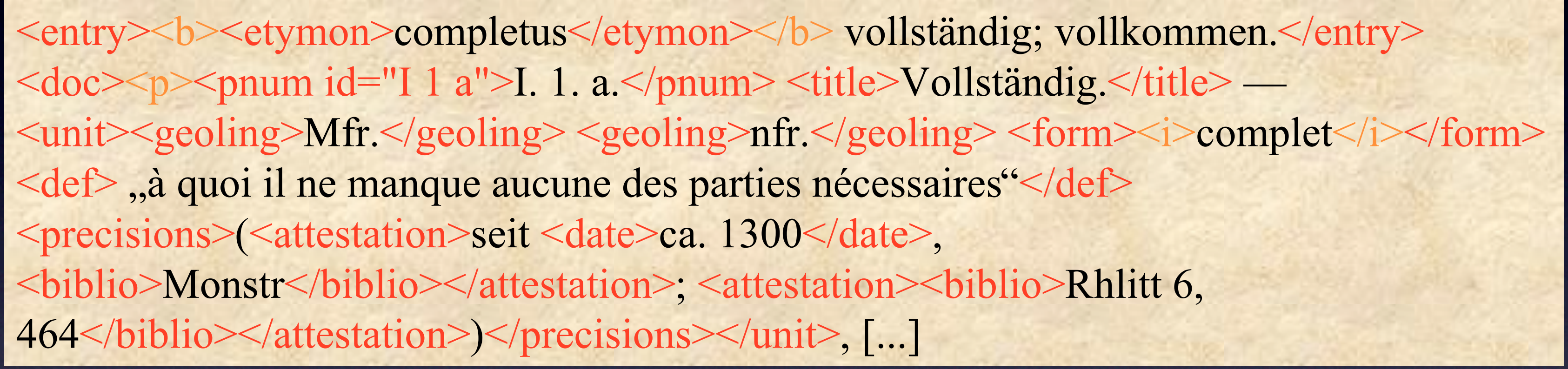


```
<b>completus</b> vollständig;<lb/> vollkommen.<lb/>
<p>I. 1. a. Vollständig. — Mfr. nfr. <i>complet</i> „à<lb/>
quoi il ne manque aucune des parties nécessaires“<lb/>
(seit ca. 1300, Monstr; Rhlitt 6, 464), saint. St-<lb/>
Seurin <i>compiet</i>, Minot <i>conpiet</i>, npr. <i>coumplèt</i>. —<lb/>
Übertragen. Nfr. <i>complet</i> „(pop.) tout à fait ivre“<lb/>
(seit Flick 1802).
```

Plain text file +

XML formatting tags : bold, italic, line break, ...

FEW Retroconversion Output



```
<entry><b><etymon>completus</etymon></b> vollständig; vollkommen.</entry>
<doc><p><pnum id="I 1 a">I. 1. a.</pnum> <title>Vollständig.</title> —
<unit><geoling>Mfr.</geoling> <geoling>nfr.</geoling> <form><i>complet</i></form>
<def> „à quoi il ne manque aucune des parties nécessaires“</def>
<precisions>(<attestation>seit <date>ca. 1300</date>,
<biblio>Monstr</biblio></attestation>; <attestation><biblio>Rhlitt 6,
464</biblio></attestation>)</precisions></unit>, [...]
```

Identical text + semantic XML tagging :

- infrastructure : <unit> + geolinguistic label, form, definition, datation, bibliographical reference, ...
- microstructure : entry / documentation / comment / notes, title, paragraph numbering, ...

A.3. Exploitation

Semantic Search

- users interested to search the contents and attributes of tags, not only the textual contents of the article

- when the retroconversion project is completed, retroconverted tagged articles will be made semantically searchable

Transversal Search

Important class of semantic search
= multicriteria search across the whole dictionary

- what vocabulary was created in the 16th century?

  depends on : <form>, <date>

- what are the French words derived from Greek?

  depends on : <form>, <lang_etymon>, <geoling>

- what is the vocabulary of a specific dialect?
- what are the words that a specific author was the first to introduce?

Enhanced Visualization

retroconversion enables to:

- resolve, independently of syntactic variations:

  4000+ geolinguistic labels (e.g. “nfr.” => français moderne, “saint.” => saintongeais)

  8000+ bibliographic labels (e.g. “Gl” => Glossaire des patois de la Suisse Romande)

- highlight the structure of the article

  with coloured text and a table of contents

Outline

A. Presentation of the Project

B. The Retroconversion System

1. Architecture
2. Algorithm Design
3. Algorithms : Complete Example
4. In Practice

B.1. Architecture

Retroconversion Workflow

To retronvert one article :

STEP 1 : digitize (+ ocerize) the paper article, including its formatting (bold, italic, paragraph/notes delimiters, volume/book/page/column/in-column numbering)

→ XML file complying with FFML Schema (formatting tags)

STEP 2 : retroconvert the article

→ XML file complying with FSML Schema (semantic tags)

Why automate the tagging of semantic concepts ?

It is important that articles be semantically tagged in a consistent manner.

- too many articles
- not enough human experts able to disambiguate the implicitness
- error-prone task

Design choice :

- automate as much as possible (100% ?)
- let human experts review hard cases that cannot be handled by our proposed automata

Retroconversion Questions

- WHAT tags should be inserted ?
  no complete model of the ***real*** FEW exists... variations variations variations

- WHERE should tags be inserted ?
  detection criteria must be reliable based on limited information

- WHEN should tags be inserted ?
  avoid interferences, e.g. tag X before tag Y ?

- HOW should tags be detected and inserted ?
  find the right software tools

Modeling the FEW (what)

The XML tagging has to

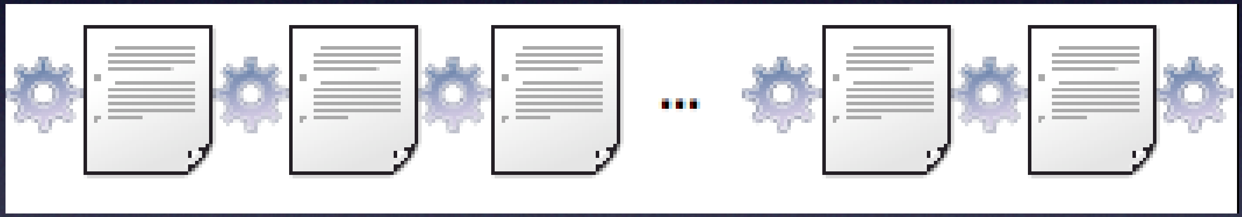
- be adapted to the structure of the dictionary
- enable semantic search

So, we have to

- create a [set of partial models, not a full] model of the structure of the FEW
- identify users' needs

Algorithm Sequence (when)

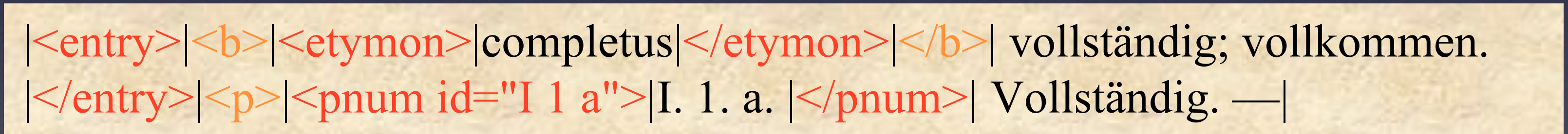
Each specific informational field is tagged by a specific algorithm.



Technology (how)

Existing XML technology intended for tree-based search and update, not for text-based search and update.

Everything's a text chunk or a tag :



```
|<entry>|<b>|<etymon>|completus|</etymon>|</b>| vollständig; vollkommen.
|</entry>|<p>|<pnum id="I 1 a">|I. 1. a. |</pnum>| Vollständig. —|
```

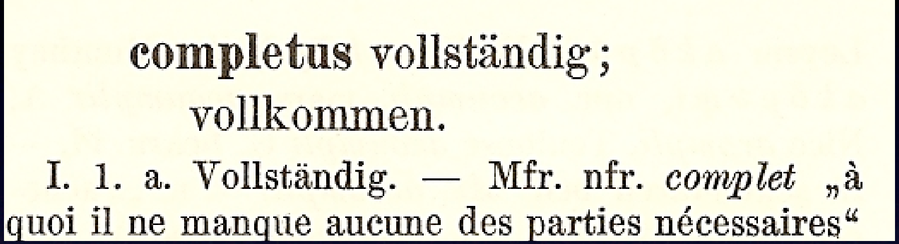
B.2. Algorithm Design

Recognition Criteria (Linguistics)

Looking into the printed version, we try to find for each information :

- typographical criteria
  - ⟶ italic, bold, small caps, specific punctuation, ...
- lexical criteria
  - ⟶ specific words
- positional criteria
  - ⟶ specific position in the structure of the FEW

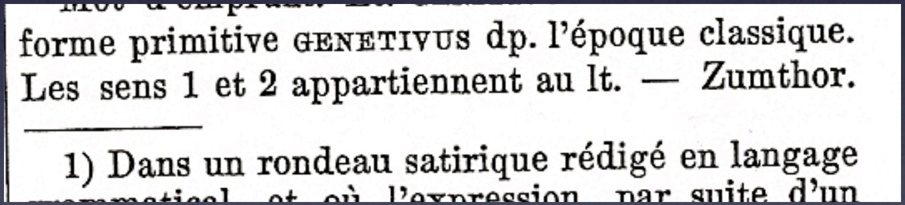
Recognition Criteria : Examples



completus vollständig;
vollkommen.

I. 1. a. Vollständig. — Mfr. nfr. *complet* „à quoi il ne manque aucune des parties nécessaires“

Etymons : specific words like “completus” (lexical), in bold (typographical) and situated at the beginning of the entry (positional)



forme primitive GENETIVUS dp. l’époque classique. Les sens 1 et 2 appartiennent au lt. — Zumthor.

1) Dans un rondeau satirique rédigé en langage grammatical, et où l’expression, par suite d’un

Signatures : specific words like “Zumthor” (lexical), situated at the end of the article (positional), preceded by — and followed by a point (typographical).

Recognition Criteria (XML files)

Looking into the XML files, algorithms detect

- keywords (e.g. “completus”, “Zumthor”)
- patterns (e.g. punctuation)
- formatting tags (e.g. <b>, <i>)
- semantic tags inserted by previous algorithms, e.g. <entry>)

Recognition Criteria : Examples

```
<entry><b>completus</b> vollständig;<lb/> vollkommen.</entry><lb/>
<p>I. 1. a. Vollständig. — Mfr. nfr. <i>complet</i> „à<lb/>
quoi il ne manque aucune des parties nécessaires“<lb/>
```

IF the first word after <entry> (semantic tag)

is “completus” (keyword)

and is surrounded by <b>... </b> (formatting tags)

THEN it has to be tagged as an <etymon>.

Algorithm Design

Methodology of algorithm design:

- select criteria (over tags, keywords, ...)
- find a combination to obtain a reliable and not ambiguous detector
- test algorithm on corpus, in the context of the retroconversion sequence
- repeat steps above until algorithm is sufficiently reliable :)

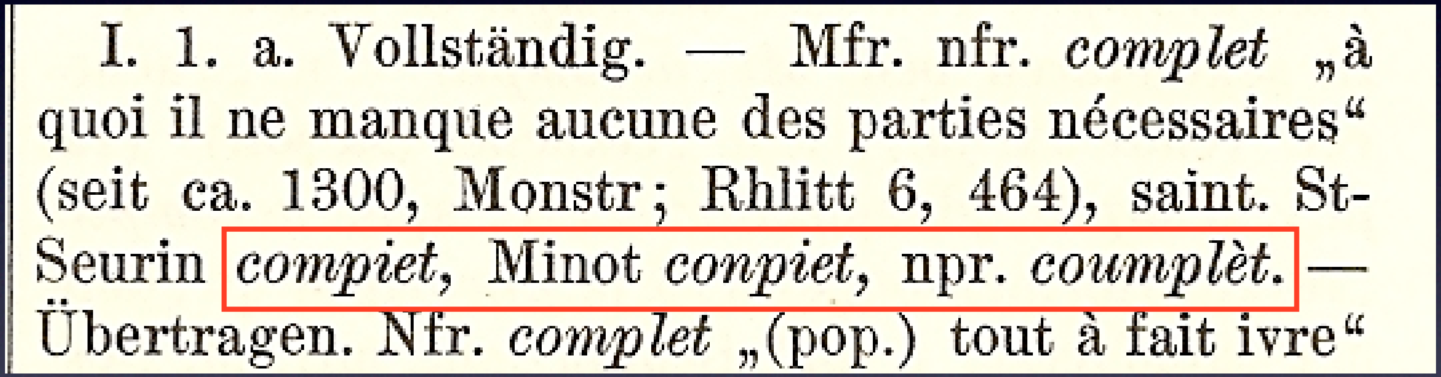
Iterative Algorithm Design

Some criteria are reliable, but would be ambiguous
because they’re also reliable for others informational fields.

Example : “Chambon” (keyword)

- signature (= Jean-Pierre Chambon) ?
- geolinguistic label (= Le Chambon-le-Château, Lozère, France) ?

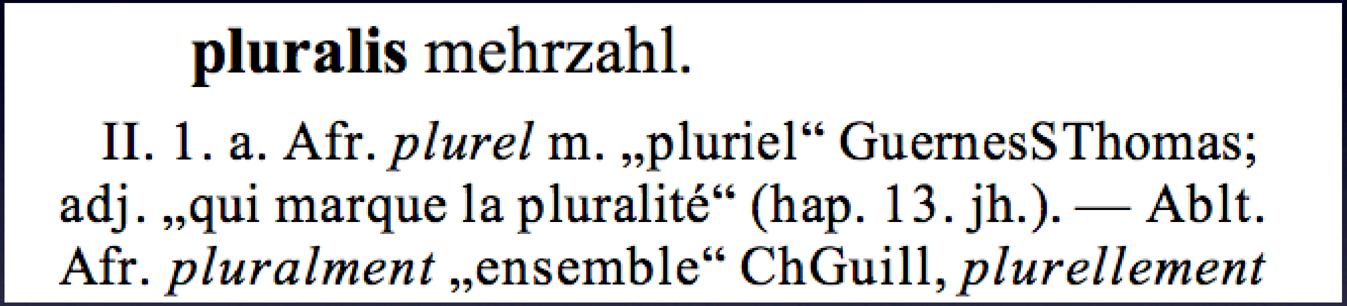
Handling of the implicitness



I. 1. a. Vollständig. — Mfr. nfr. *complet* „à quoi il ne manque aucune des parties nécessaires“ (seit ca. 1300, Monstr; Rhlitt 6, 464), saint. St-Seurin *compiet*, Minot *conpiet*, npr. *coumplèt*. — Übertragen. Nfr. *complet* „(pop.) tout à fait ivre“

Example : Implicit Definitions

Inconsistencies handling



pluralis mehrzahl.

II. 1. a. Afr. *plurel* m. „pluriel“ GuernesSThomas; adj. „qui marque la pluralité“ (hap. 13. jh.). — Ablt. Afr. *pluralment* „ensemble“ ChGuill, *plurellement*

Example : wrong paragraph numbering

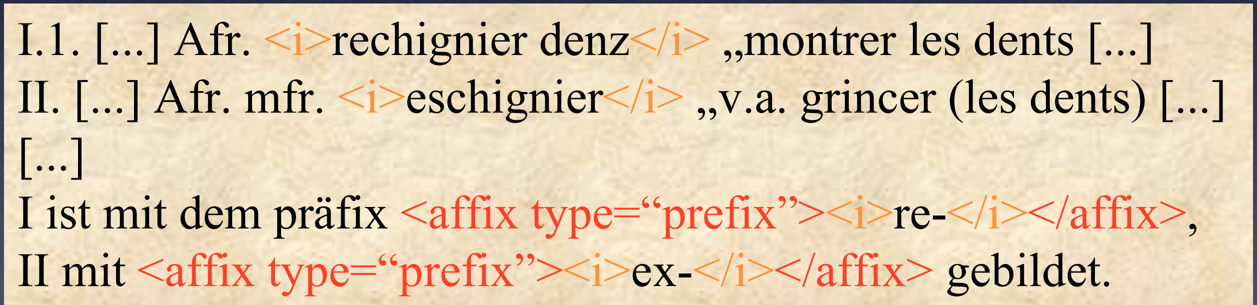
The FEW was written from 1922 to 2005 by several people, and thus contains a lot of mistakes or inconsistencies.

B.3. Algorithms : Complete Example

Tagging of Affixes

affix = morpheme that can be prepended/appended to a word to form a new word

Example (FEW 16, 323a, *KINAN) :



I.1. [...] Afr. <i>rechignier denz</i> „montrer les dents [...]
II. [...] Afr. mfr. <i>eschignier</i> „v.a. grincer (les dents) [...]
[...]
I ist mit dem präfix <affix type="prefix"><i>re-</i></affix>,
II mit <affix type="prefix"><i>ex-</i></affix> gebildet.

Article Processing

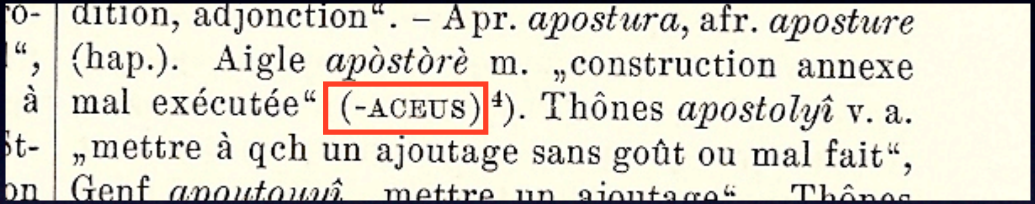
for each paragraph :

- tag suffixes
- then tag prefixes

for each paragraph :

- tag French affixes

Suffix Tagging



dition, adjonction“. – Apr. *apostura*, afr. *aposture* (hap.). Aigle *apòstòrè* m. „construction annexe mal exécutée“ (-ACEUS) [4]). Thônes *apostolyî* v. a. „mettre à qch un ajoutage sans goût ou mal fait“, Genf *apoutouyî* „mettre un ajoutage“ Thônes

search the paragraph's text for all keywords
from suffix keyword list (e.g. “-abundu”, “-aga”, “-amen”, ...)

for each found suffix keyword :

- check if it can be extended to the left
- tag it

Prefix Tagging

search the paragraph's text for all keywords
from prefix keyword list (e.g. ab-, ad-, archi-, bene-,...)

- for each found prefix keyword :
    - filter out keywords followed by a line break
    - check if it can be extended to the right
    - tag it

Surprise...

Line breaks can appear everywhere !

Example :

the XML input file contains : ex-<lb />tra-

the algorithm should see : extra-

→ dash-aware keyword search

Surprise...

- Some keywords can appear in definitions, etymons, ...
- Note references can appear everywhere

Prior to searching for prefixes and suffixes, make invisible the definitions, etymons, exponents, note references, ...

→ tag-aware keyword search

French Affixes Tagging

for each candidate <i>...</i>* :

- check left and right contexts for hint

  (i.e. at most 10** words away from the candidate,
  find one of “préfix”, “suffix”, “affix”, “confix”, “ablt.”, “ableit”,
  e.g. “suffix bla bla bla <i>-illon</i>”)

* candidate = 2+ characters, starts and/or ends with a dash

** 10 = arbitrary choice, we can only do heuristics, not optimal algorithms

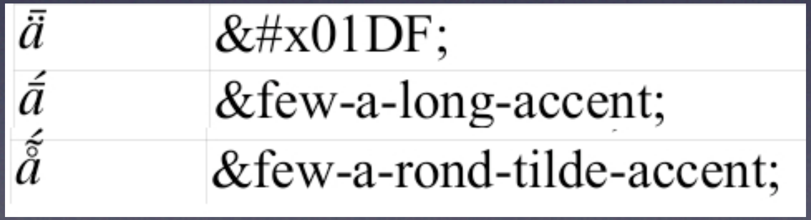
B.4. In Practice

Character Coding

- When you type the letter F, E then W on your keyboard, what is typically stored on your computer ?
(answers: the numbers 70, 69 and 87)

- Computers store numeric codes only ;
each of these numeric codes shoud be mapped to a glyph,
i.e. symbol that is printed on screen

- What about these?



ǟ	&#x01DF;
ā́	&few-a-long-accent;
å̃́	&few-a-rond-tilde-accent;

Unicode UTF-8 character coding

- Unicode = computer standard to code
  more than 100 000 characters from many languages
- UTF-8 = most widespread and well-supported
  character coding able to represent Unicode characters
- More than 125 characters found in the FEW
  yet to be included into Unicode...

→ use Unicode's so-called "private zone"
until these characters are included into Unicode

Systematic Coding

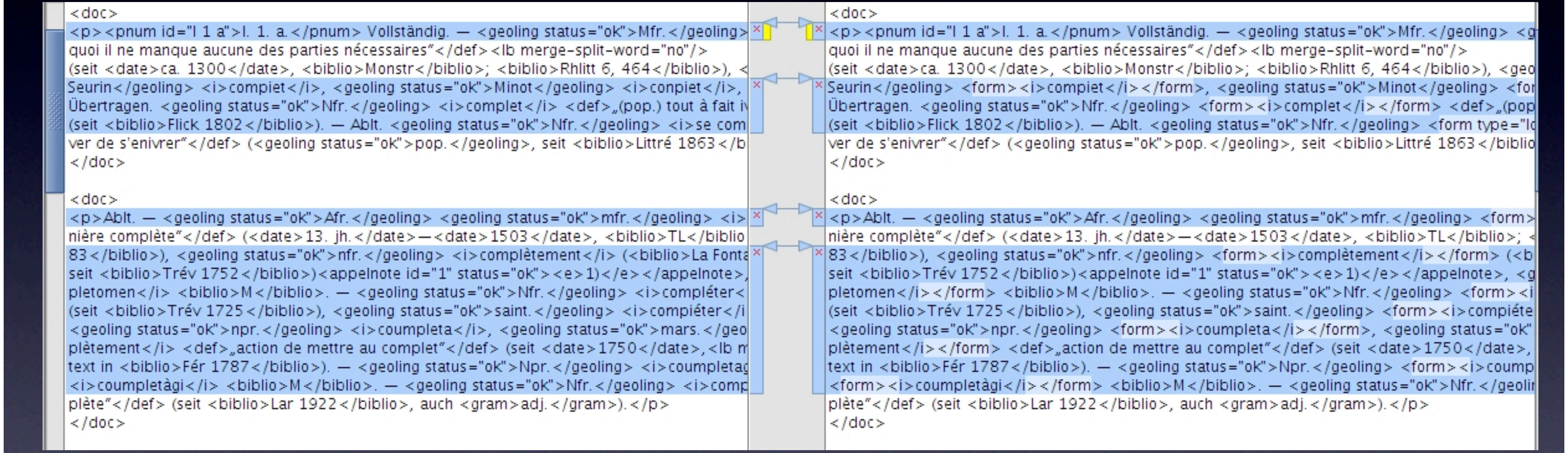
- Both XML files and keyword lists must use
  the same "enhanced" UTF-8 coding
  (i.e. UTF-8 with 125+ special FEW characters)
- ...else keyword search will fail,
  i.e. will not find keywords featuring special characters!
- Do pay attention to the coding of your data

Retroconversion History

- Requirement to enable human experts
  to review the behaviour of the retroconversion software
  and to check every character inserted/updated into the XML files

- All ~35 intermediate XML files
  resulting from the retroconversion of 1 article
  together constitute its "historical log"

Retroconversion History



```
<doc>
<p><pnum id="l 1 a">I. 1. a.</pnum> Vollständig. — <geoling status="ok">Mfr.</geoling>
quoi il ne manque aucune des parties nécessaires"</def><lb merge-split-word="no"/>
(seit <date>ca. 1300</date>, <biblio>Monstr</biblio>; <biblio>Rhlitt 6, 464</biblio>), <
Seurin</geoling> <i>compiet</i>, <geoling status="ok">Minot</geoling> <i>conpiet</i>,
Übertragen. <geoling status="ok">Nfr.</geoling> <i>complet</i> <def>„(pop.) tout à fait iv
(seit <biblio>Flick 1802</biblio>). — Ablt. <geoling status="ok">Nfr.</geoling> <i>se com
ver de s'enivrer"</def> (<geoling status="ok">pop.</geoling>, seit <biblio>Littré 1863</b
</doc>

<doc>
<p>Ablt. — <geoling status="ok">Afr.</geoling> <geoling status="ok">mfr.</geoling> <i>
nière complète"</def> (<date>13. jh.</date>—<date>1503</date>, <biblio>TL</biblio
83</biblio>), <geoling status="ok">nfr.</geoling> <i>complètement</i> (<biblio>La Fonta
seit <biblio>Trév 1752</biblio>)<appelnote id="1" status="ok"><e>1)</e></appelnote>,
pletomen</i> <biblio>M</biblio>. — <geoling status="ok">Nfr.</geoling> <i>compléter<
(seit <biblio>Trév 1725</biblio>), <geoling status="ok">saint.</geoling> <i>compiéter</i
<geoling status="ok">npr.</geoling> <i>coumpleta</i>, <geoling status="ok">mars.</geo
plètement</i> <def>„action de mettre au complet"</def> (seit <date>1750</date>,<lb m
text in <biblio>Fér 1787</biblio>). — <geoling status="ok">Npr.</geoling> <i>coumpletag
<i>coumpletàgi</i> <biblio>M</biblio>. — <geoling status="ok">Nfr.</geoling> <i>comp
plète"</def> (seit <biblio>Lar 1922</biblio>, auch <gram>adj.</gram>).</p>
</doc>
```

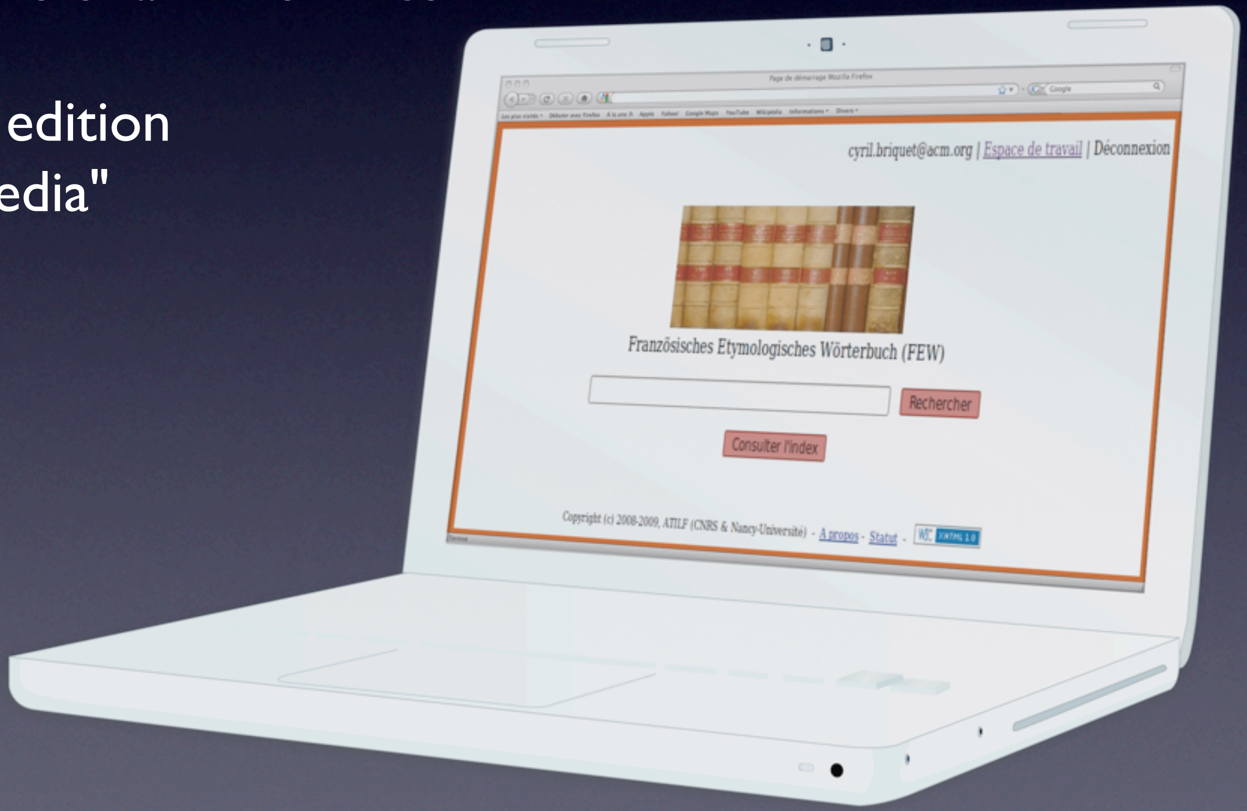
```
<doc>
<p><pnum id="l 1 a">I. 1. a.</pnum> Vollständig. — <geoling status="ok">Mfr.</geoling> <g
quoi il ne manque aucune des parties nécessaires"</def><lb merge-split-word="no"/>
(seit <date>ca. 1300</date>, <biblio>Monstr</biblio>; <biblio>Rhlitt 6, 464</biblio>), <geo
Seurin</geoling> <form><i>compiet</i></form>, <geoling status="ok">Minot</geoling> <for
Übertragen. <geoling status="ok">Nfr.</geoling> <form><i>complet</i></form> <def>„(pop
(seit <biblio>Flick 1802</biblio>). — Ablt. <geoling status="ok">Nfr.</geoling> <form type="lo
ver de s'enivrer"</def> (<geoling status="ok">pop.</geoling>, seit <biblio>Littré 1863</biblio
</doc>

<doc>
<p>Ablt. — <geoling status="ok">Afr.</geoling> <geoling status="ok">mfr.</geoling> <form>
nière complète"</def> (<date>13. jh.</date>—<date>1503</date>, <biblio>TL</biblio>; <
83</biblio>), <geoling status="ok">nfr.</geoling> <form><i>complètement</i></form> (<b
seit <biblio>Trév 1752</biblio>)<appelnote id="1" status="ok"><e>1)</e></appelnote>, <g
pletomen</i></form> <biblio>M</biblio>. — <geoling status="ok">Nfr.</geoling> <form><i
(seit <biblio>Trév 1725</biblio>), <geoling status="ok">saint.</geoling> <form><i>compiéte
<geoling status="ok">npr.</geoling> <form><i>coumpleta</i></form>, <geoling status="ok"
plètement</i></form> <def>„action de mettre au complet"</def> (seit <date>1750</date>,
text in <biblio>Fér 1787</biblio>). — <geoling status="ok">Npr.</geoling> <form><i>coump
<form><i>coumpletàgi</i></form> <biblio>M</biblio>. — <geoling status="ok">Nfr.</geolin
plète"</def> (seit <biblio>Lar 1922</biblio>, auch <gram>adj.</gram>).</p>
</doc>
```

Web-based Retroconversion Platform (prototype)

20 000 articles in the FEW x ~35 XML files per article
= ~ three quarters of a million files

⟶ distributed edition
"à la Wikipedia"



Page de démarrage Mozilla Firefox

cyril.briquet@acm.org | Espace de travail | Déconnexion

Französisches Etymologisches Wörterbuch (FEW)

Copyright (c) 2008-2009, ATILF (CNRS & Nancy-Université) - A propos - Statut -

Cooperation Between Linguistics and Software Engineering

What was difficult?

- Linguistics part :
  providing accurate model of the *real* FEW *in a timely fashion*
  (if too simple: many fields not properly tagged ;
  if too complex: too costly to integrate into the software)

- Software engineering part :
  adapting to *varying specifications* following *experiments on real articles*

Cooperation Between Linguistics and Software Engineering

What was key to success of the project?

- Linguistics part :
  finding detection criteria in function of
  feedback from experiments, time available,
  and available software support (and shortcomings :-)

- Software engineering part :
  providing *dedicated* and *flexible* software tools
  to support specialized and complex linguistic reasoning

Theory and Practice

Theory and practice should be synchronized.

"In theory, there's no difference between theory and practice.
In practice, there is." *Yogi Berra*

Outline

A. Presentation of the Project

B. The Retroconversion System

C. Beyond Retroconversion

Beyond Retroconversion

When the retroconversion project is completed, the next step will be to make the tagged articles semantically searchable

→ i.e. search the contents and attributes of tags, not only in the textual contents of the article

To achieve acceptable performance: dedicated search engine

- ➢ requires to index the articles
- ➢ requires to specify which queries are expected

Beyond Retroconversion

Exploitation of the retroconverted dictionary :

- easier reading
- transversal (semantic) search
- available from a website

- updates
- links with other dictionaries (DEAF, TLF, ...)
- links with other FEW projects

Conclusion

Conclusion

Before starting the project, necessity of knowing

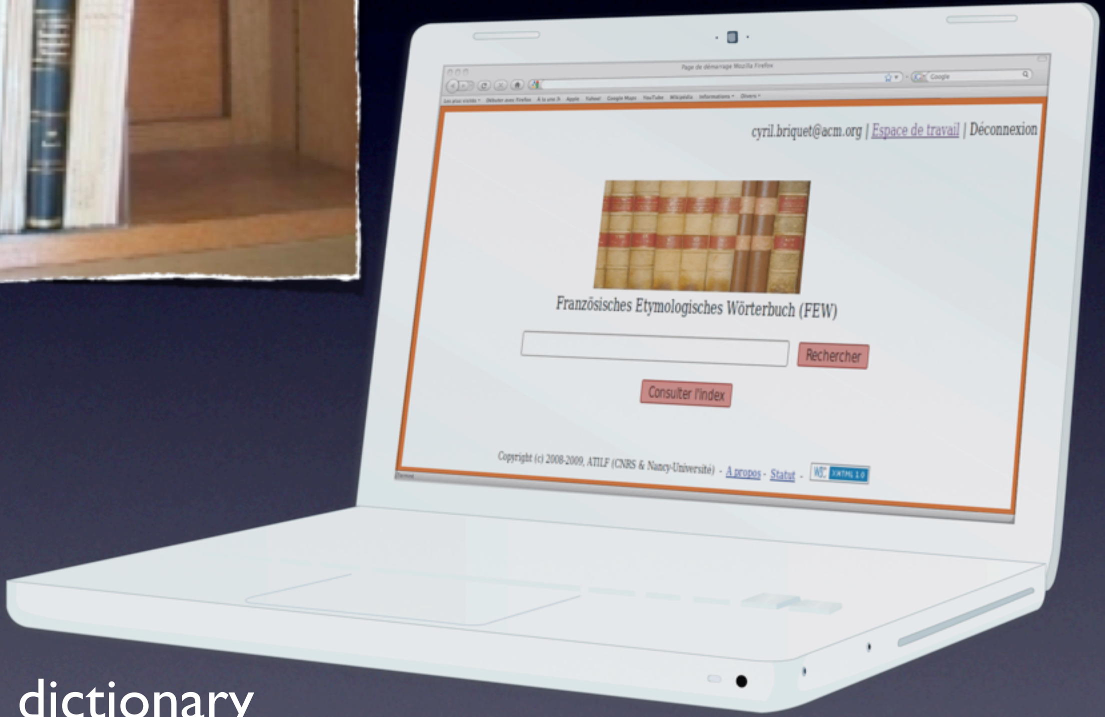
- the structures of the dictionary :

  metalexicographic study, e.g. Büchi 1996

- the users' needs

During the project, necessity to

- iterate on the algorithms and on the dictionary model



cyril.briquet@acm.org | Espace de travail | Déconnexion

Französisches Etymologisches Wörterbuch (FEW)

Copyright (c) 2008-2009, ATILF (CNRS & Nancy-Université) - A propos - Statut -

Having a retroconverted dictionary
opens up exciting new possibilities to users,
in particular for such a complex dictionary that is not easily accessible

Bibliography

Büchi, E., 1996. *Les Structures du* Französisches Etymologisches Wörterbuch. *Recherches métalexicographiques et métalexicologiques*. Tübingen.

DEAF = Baldinger, K. *et al*., 1971–. *Dictionnaire étymologique de l’ancien français.* Québec/Tübingen/Paris.

FEW = Wartburg, W. von *et al.* (1922-2002). *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes*. 25 vol. Bonn/Heidelberg/Leipzig-Berlin/Bâle.

LEI = Pfister, M./Schweickard, W. (dir.), 1979–. *Lessico etimologico italiano*. Wiesbaden.

TLF = Imbs, P. (dir.), 1971–1994. *Trésor de la langue française. Dictionnaire de la langue du XIX*^{e} *et du XX*^{e} *siècle (1789-1960)*. 16 vol. Paris.