Balisage
The Markup Conference
2010

A Virtualization-Based Retrieval and Update API for XML-Encoded Corpora

Cyril Briquet (1) (2), Pascale Renders (2) (3), Etienne Petitjean (2)

(1) McMaster U, ON, Canada (2) CNRS, Nancy, France (3) U of Liège, Belgium

Take-home message

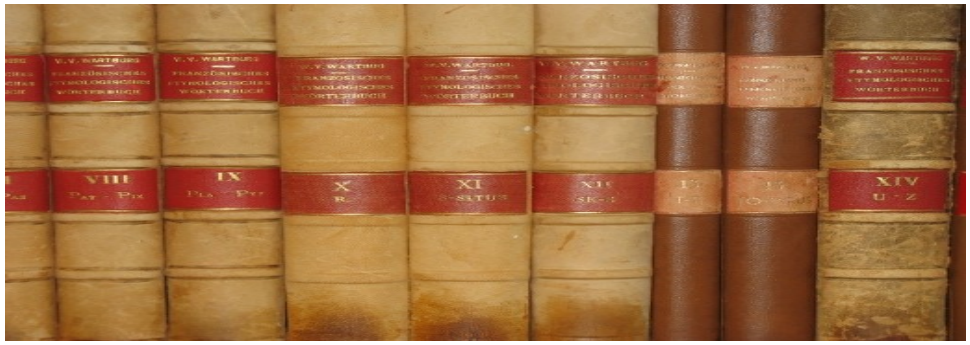
- context: FEW, ref. dictionary in French & Romance Linguistics
- objective: semantic tagging of a very very complex dictionary
- our desire: offer support for *natural linguistic reasoning*
  = tag-aware text retrieval, tag-aware markup update
- our proposed mechanism (made available as an API):
  virtualizing sections of the XML document as needed
- <disclaimer>we're not XML experts</disclaimer> <!-- ;) -->

This afternoon's agenda

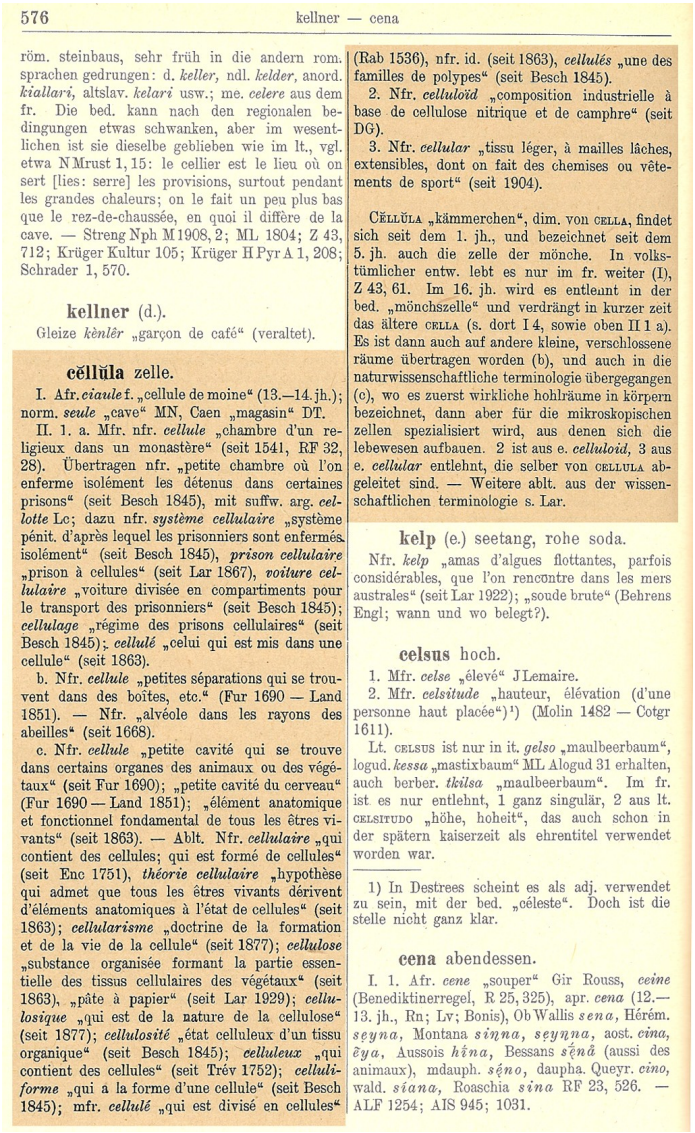
- FEW dictionary
- the retroconversion problem
- virtualizing the XML document (concept, API)
- in practice

Französisches Etymologisches Wörterbuch

- reference dictionary
  in French & Romance Linguistics
- Walther von Wartburg et al.,
  1922-2002



- historical & etymological



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 bedingungen etwas schwanken, aber im wesentlichen ist sie dieselbe geblieben wie im lt., vgl. etwa N Mrust 1, 15: le cellier est le lieu où on sert [lies: serre] les provisions, surtout pendant les grandes chaleurs; on le fait un peu plus bas que le rez-de-chaussée, en quoi il diffère de la cave. — Streng Nph M 1908, 2; ML 1804; Z 43, 712; Krüger Kultur 105; Krüger H Pyr A 1, 208; Schrader 1, 570.

kellner (d.).
Gleize *kènlêr* „garçon de café“ (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 1867), *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 qui admet que tous les êtres vivants dérivent d'éléments anatomiques à l'état de cellules“ (seit 1863); *cellularisme* „doctrine de la formation et de la vie de la cellule“ (seit 1877); *cellulose* „substance organisée formant la partie essentielle des tissus cellulaires des végétaux“ (seit 1863), „pâte à papier“ (seit Lar 1929); *cellulosique* „qui est de la nature de la cellulose“ (seit 1877); *cellulosité* „état celluleux d'un tissu organique“ (seit Besch 1845); *celluleux* „qui contient des cellules“ (seit Trév 1752); *celluliforme* „qui a la forme d'une cellule“ (seit Besch 1845); mfr. *cellulé* „qui est divisé en cellules“ (Rab 1536), nfr. id. (seit 1863), *cellulés* „une des familles de polypes“ (seit Besch 1845).
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 „kämmerchen“, dim. von cella, findet sich seit dem 1. jh., und bezeichnet seit dem 5. jh. auch die zelle der mönche. In volkstümlicher entw. lebt es nur im fr. weiter (I), Z 43, 61. Im 16. jh. wird es entlehnt 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 zellen spezialisiert wird, aus denen sich die lebewesen aufbauen. 2 ist aus e. *celluloïd*, 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“ J Lemaire.
2. Mfr. *celsitude* „hauteur, élévation (d'une personne haut placée“)[1]) (Molin 1482 — Cotgr 1611).
Lt. celsus ist nur in it. *gelso* „maulbeerbaum“, logud. *kessa* „mastixbaum“ ML Alogud 31 erhalten, auch berber. *tkilsa* „maulbeerbaum“. Im fr. ist es nur entlehnt, 1 ganz singulär, 2 aus lt. celsitudo „höhe, hoheit“, das auch schon in der spätern kaiserzeit als ehrentitel verwendet 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. *seyna*, Montana *sinna*, *seynna*, aost. *cina*, *ẽya*, Aussois *hĩna*, Bessans *sẽnâ* (aussi des animaux), mdauph. *séno*, daupha. Queyr. *cino*, wald. *siana*, Roaschia *sina* RF 23, 526. — ALF 1254; AIS 945; 1031.

Shallow comparison: OED & FEW

Feature	OED	FEW
Pages	21730	16865
Volumes	20	25
Entries	300 000	20 000 (*)
Lexemes	600 000	900 000 (est.)

(*) FEW entries are etymons, not lexemes, thus fewer

FEW is very very complex

hard to read:

- complex structure
- large number of fields
- implicitness (syntactic + semantic)

hard to search:

- can't do transversal search in paper version

Retroconversion of the FEW

<< starting from the paper version,

how can the complex dictionary structure

be automatically extracted into a searchable database? >>

atilf ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

* ongoing project at ATILF lab in Nancy, France
* team of Prof. Eva Buchi, Research Director
* backed by CNRS and Nancy University

The bottom line: an example

IN

```
<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).
```

OUT

```
<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<lb/>quoi il ne manque aucune des parties
nécessaires“</def><lb/>
<precisions>(<attestation>seit <date>ca. 1300</date>,
<biblio>Monstr</biblio></attestation>; <attestation><biblio>Rhlitt 6,
464</biblio></attestation>)</precisions></unit>, [...]
```

Text-oriented XML documents

FEW article

=

text-oriented XML document,

complying with XML Schema

(currently not TEI but long term it'll try & align with TEI)

=

list of text chunks with interspersed tags

(element hierarchy useless, thus not used)

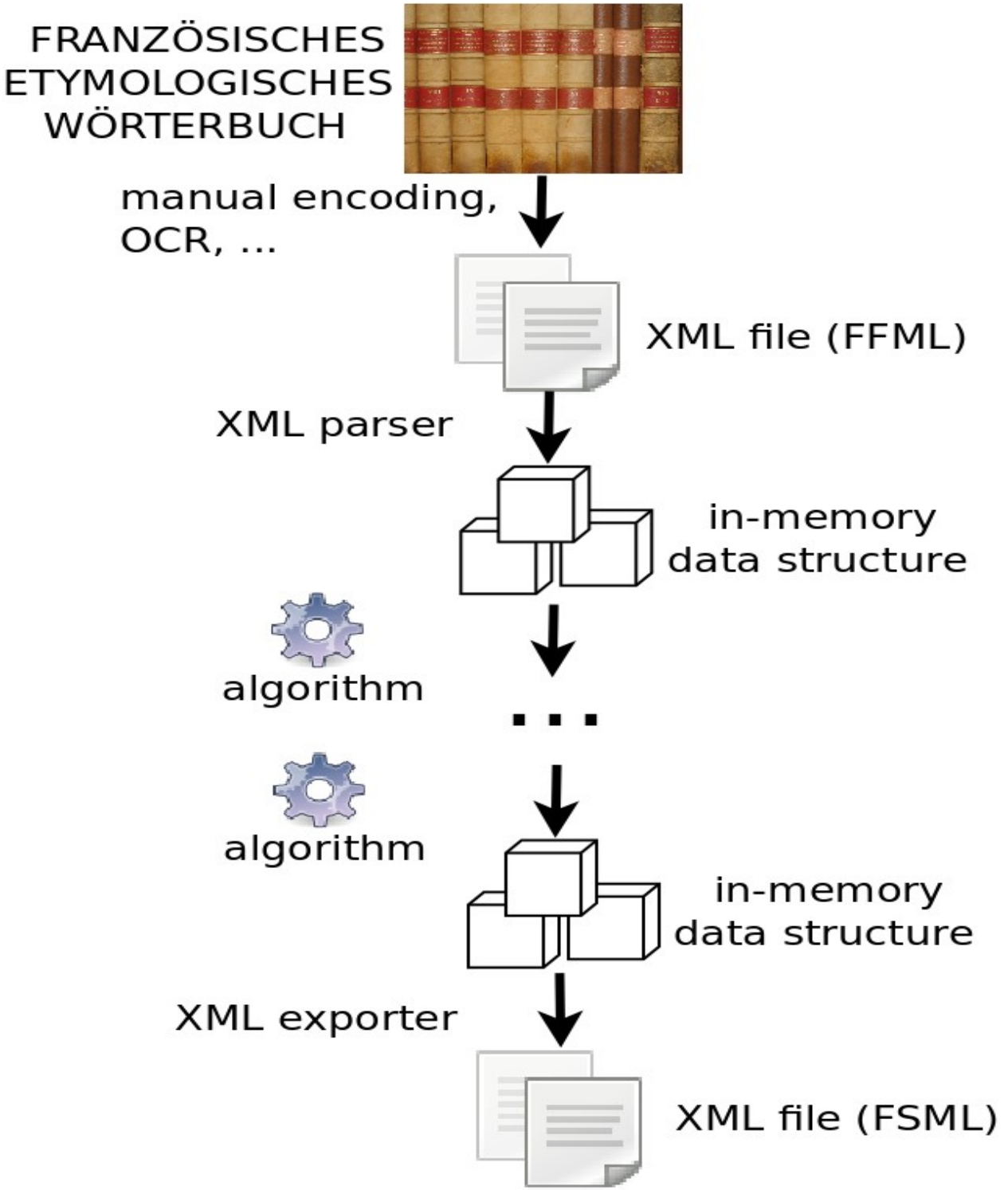
In-memory data structure

- list of nodes: XML tags or text chunks
- constructed using a validating SAX parser
- UTF-8, entities resolved, character legality enforced
- text normalized (redundant spacing, break tags)

FEW

retroconversion

workflow



What's in a tagging algorithm?

- detection of dictionary fields
  - text retrieval, markup retrieval
  - keyword search (*dictionary-matching problem*)
  - regexp
  - secondary contextual lookups often necessary,

    e.g. find keywords within 10 words of tags containing keyword,

    in text-oriented representation

- tagging of detected fields (markup update)

- sometimes, modification of dictionary text (text update)

Retrieval challenges

- false negatives:

  tag interference (e.g. exponent, end of line)

  prevents matching of keywords, regexp

- false positives in irrelevant contexts:

  keyword search not relevant everywhere

Use case: preventing false negatives

- <p>Emprunt de <geoling>Ittard.</geoling> <geoling>mlt.</geoling> <etymon>augmentator</etymon> (**4<e>e</e>–<lb/>6<e>e</e> s.**, <biblio>ThesLL</biblio> ;

- in this use case: **4<e>e</e>–<lb/>6<e>e</e> s.** is a datation;

  full-text query not discarding tags would result in false negative,

  as none of the 6 fragments (**4**, **e**, **-**, **6**, **e**, **s.**) alone is a datation

- in this use case: <e> tags should be skipped

  Emprunt de Ittard. mlt. augmentator (4e– 6e s., ThesLL ;

Use case: preventing false positives

- <geoling>Nfr.</geoling> com-<lb />plètement <def>„action de mettre au complet“</def> (seit **1750**,<lb/>text in <biblio>Fér **1787**</biblio>).

- in this use case: **1750** is a date, **1787** is not;

  full-text query only discarding all tags would result in false positive

- in this use case: <biblio> elements should be made invisible

  Nfr. complètement „action de mettre au complet“ (seit 1750, text in)

Update challenges

- updates may be far from matches,
  i.e. in non-collateral branch of tree representation

- updates may span several text chunks,
  with interferences from legitimate tags in-between

- match points required to offer support for
  natural linguistic reasoning

Virtual string

- **Definition:** *concatenation of adjacent text chunks, except those within elements configured to be invisible*

- sections of XML document virtualized into multiple virtual strings separated by visible tags

- backed by underlying XML document; updates are transparently propagated

Text virtualization example

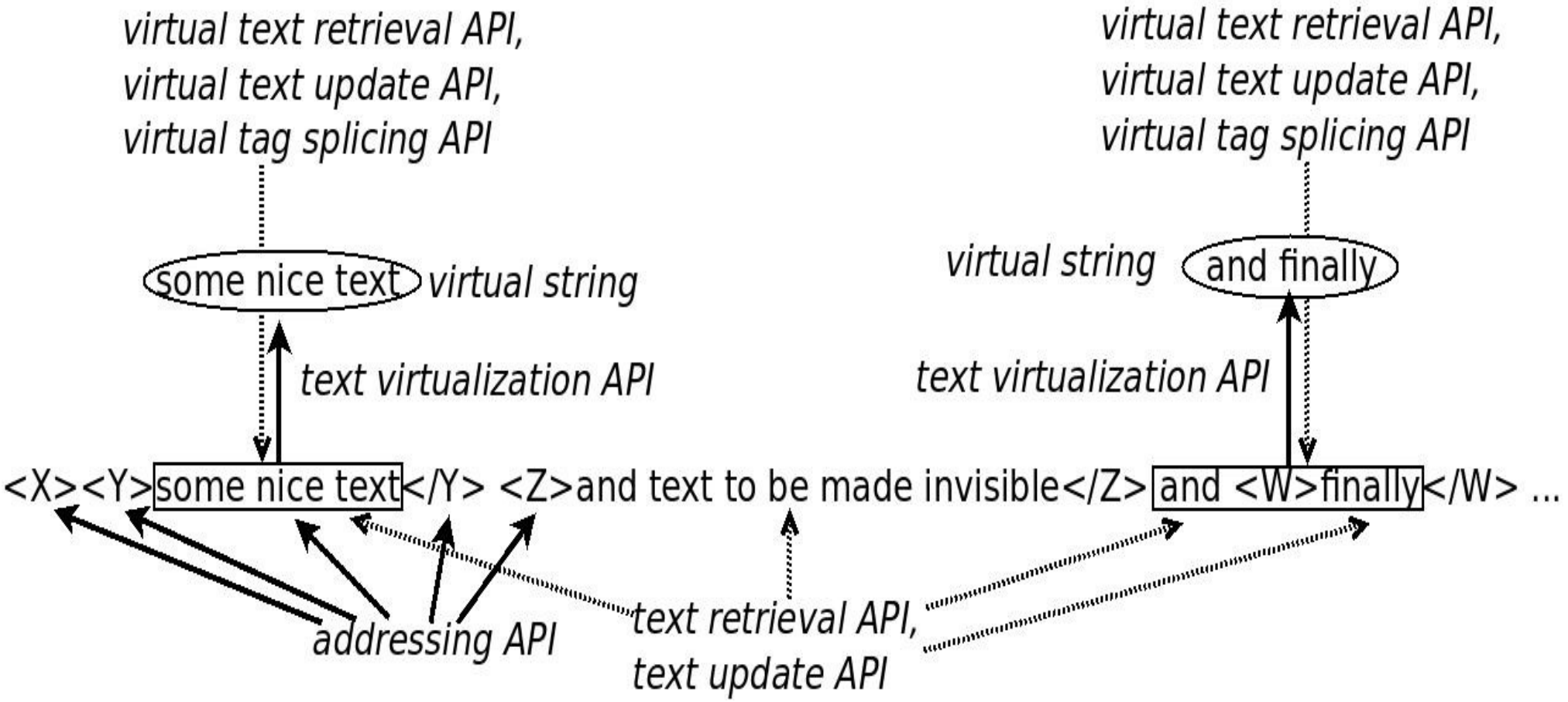
visibility: V visible, I invisible, S skipped, T terminal

3 virtual strings, tag last 2 words of middle virtual string :

- … <V>**some nice text**</V> <I>and text to be made invisible</I> **and now** <S>**finally**</S> <V>**nice text again**</V></T> ...

- … <V>some nice text</V> <I>and text to be made invisible</I> now **<NEW>**now <S>finally</S>**</NEW>** <V>nice text again</V></T> ...

API overview



read this slide bottom-up, please :-)

Syntax example

```
VirtualTextSearcher searcher = new VirtualTextSearcher(iterator, partition);
for (VirtualString vs : searcher) { // text virtualization
    Set<KeywordMatch> matches = fewPrefixBase.findAllKeywords(vs.getText());
    VirtualTagSplicer virtualTagSplicer = createVirtualTagSplicer(this,vs);
    for (KeywordMatch m : matches) {
      int startIndex = ...; int endIndex = ...;                  // virtual text retrieval:
      if (isLicitPrefix(vs,endIndex) == false) continue; // requires match point
      endIndex = getExtendedPrefixKeywordEndIndex(vs,endIndex);
      virtualTagSplicer.markSubstringForTagging(startIndex,endIndex,affix,
      new String[] { "type", "descendance" },new String[] { "prefix", "etymon" });
    }
    virtualTagSplicer.spliceAll(); // virtual tag splicing
}
```

Natural linguistic reasoning

- retroconversion of FEW = **breakthrough**
  - familiar level of abstraction: text without tags
  - flexible specification of retrieval & updates

- similar projects
  - abstraction level too far from dict.: tags everywhere
  - hard to specify: long regexp containing tags

In practice

- Java implementation: 64kloc (API core: 7.5kloc)
- 144 articles retroconverted (~0.75% of FEW)
- coverage: 98.5% automatically tagged
- precision and recall of tagging:
  - depend on accuracy of linguistic analysis, not on API (which returns exact results)
  - difficult to measure, takes days to tag manually

What about XQuery?

- XQuery Full Text extension:

  FTIgnore option configures tag visibility during search

- XQuery Update Facility

- returned results = XML elements...

  not text with support for match points...

  but at this point the tagging algorithm is just getting started

  => how to perform additional contextual search & updates ***?***

  (we just don't know...)

Next steps

- package API into dedicated library
- get feedback on syntax, semantics

  (to what extent does the API overlap with and/or benefit from and/or contribute to existing related technology?)
- optimizing current implementation for
  - speed: addressing, virtual text upd., virtual splicing
  - memory usage: text virtualization

Take-home message

- context: FEW, ref. dictionary in French & Romance Linguistics
- objective: semantic tagging of a very very complex dictionary
- our desire: offer support for *natural linguistic reasoning*
  = tag-aware text retrieval, tag-aware markup update
- our proposed mechanism (made available as an API):
  virtualizing sections of the XML document as needed
- <disclaimer>we're not XML experts</disclaimer> <!-- ;) -->

Thank you