Le Petit Larousse Illustré de 1905 en ligne : Présentation et secrets de fabrication

Hélène Manuélian - Audrey Bruscand - Nicole Cholewka - Anne-Marie Hetzel

LDI Cergy-Pontoise

UMR 7187 - CNRS - UCP - UP13

Introduction : naissance du projet

- Souhait de proposer une version informatisée du premier Petit Larousse Illustré
- Permettre sa conservation et sa diffusion (le fac-simile n'était pas encore paru)
- Permettre la consultation analogique

Buts et contraintes

- Conserver la caractéristique la plus visible du PLI : les images
- Permettre une consultation plein texte et une consultation experte
- Assurer une consultation universelle et pérenne

Une seule solution

- Une annotation sémantique du texte permettant le repérage de chaque composante des articles du dictionnaire respectant des standards de balisage informatique

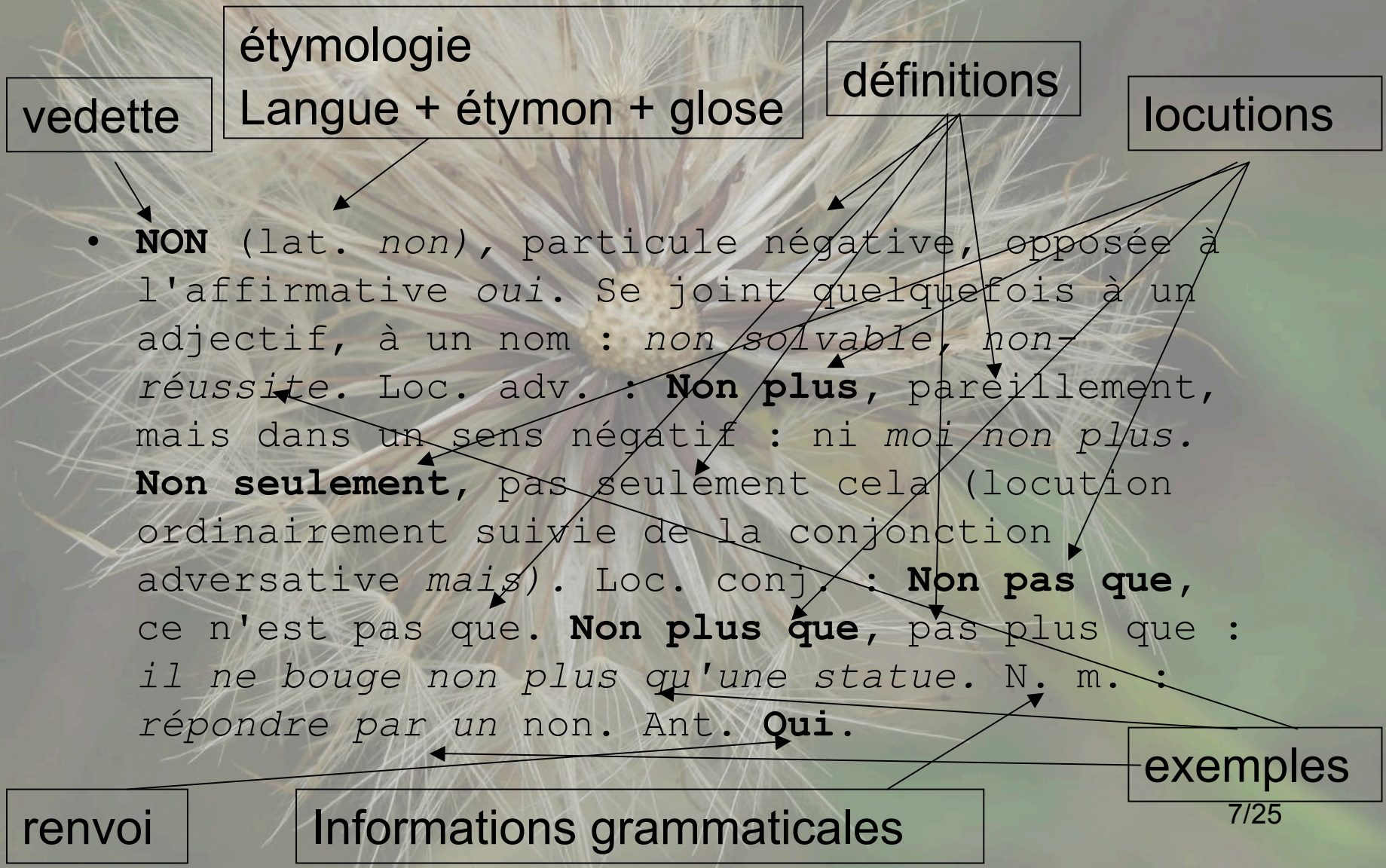
Choix en conséquence

- Une analyse fine du texte permettant un découpage précis des articles
- Une classification fine des éléments composant les articles
- Une utilisation rigoureuse (mais imaginative) de la DTD fournie par la Text Encoding Initiative (TEI) pour les dictionnaires (pour baliser le texte dans un format XML).

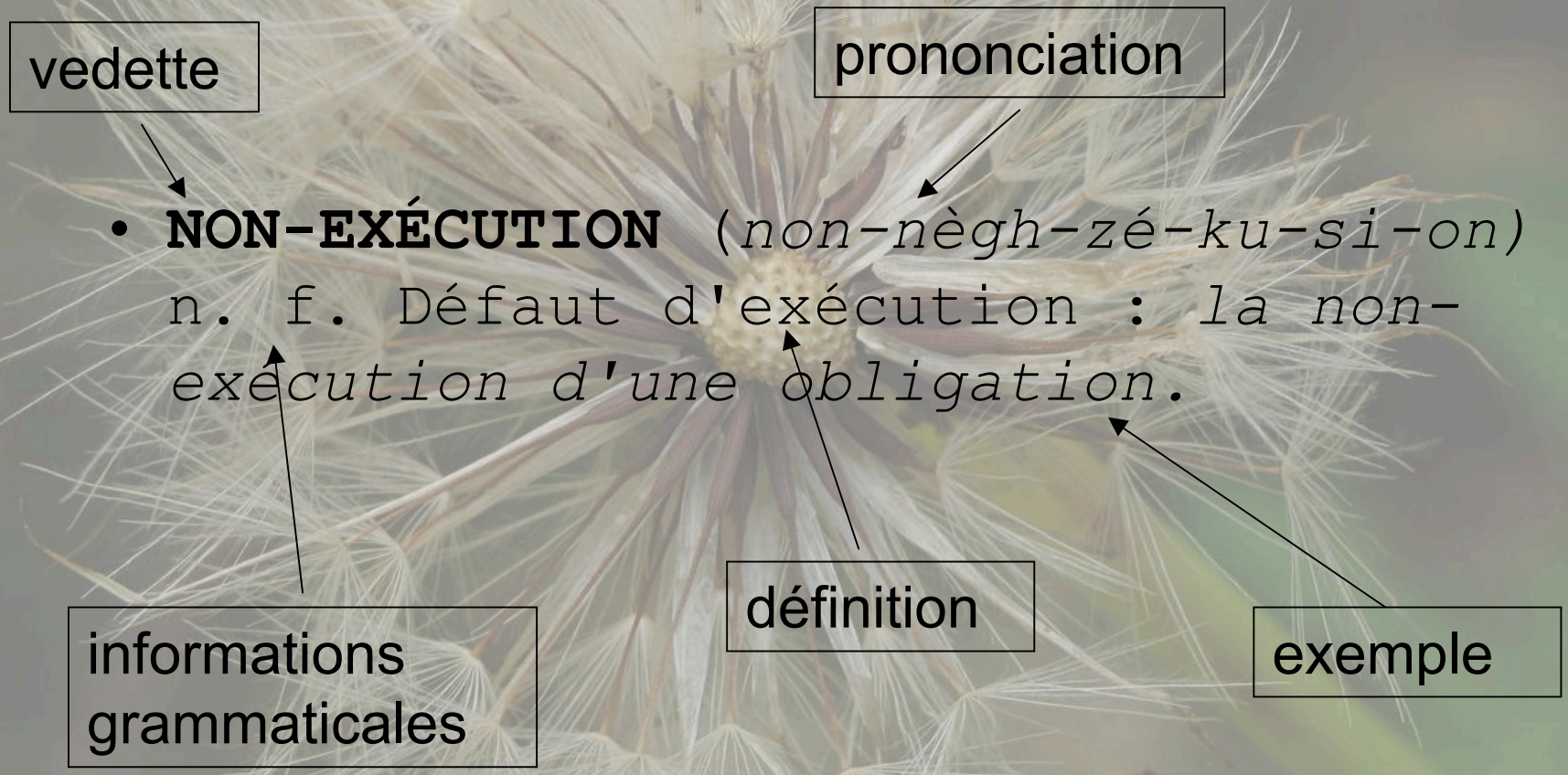
Un constat crucial

- La mise en forme du texte est très régulière : la forme « physique » du texte va pouvoir servir au balisage « sémantique » du texte

Mise en forme et sémantique



Mise en forme et sémantique



Etapes de la fabrication -1

- Numérisation (Scanner, reconnaissance optique de caractères)
- Relecture et correction du texte (humaine)
- Conversion en HTML
- Informations sur la **forme** des articles
- A_relu.html

Etapes de la fabrication -2 Première passe d'annotation

- Repérage des éléments les plus évidents :
- La vedette : balise <orth> de la TEI - quand les vedettes sont complexes (quand elles incluent une forme fléchie, le balisage reconstitue le mot fléchi, de façon à permettre la recherche sur la forme fléchie)

```
<EntryFree><form><orth>APPROXIMATIF</orth></form>,
  <form type="infl"><orth type="partial"
  mot_entier="approximative">IVE</orth> <pron>(a-pro-
  ksi)</pron></form>
  <gramGrp><pos>adj.</pos></gramGrp>
```

Etapes de la fabrication -2
Première passe d'annotation

- La prononciation : balise <pron>
- L'ensemble (vedette et prononciation) forme un élément appelé <form>
- Les informations grammaticales (<gramGrp> regroupe <pos>,<gen> et <number>)
- Les informations sur l'usage des mots (registre, style, etc.) : <usg>
- Ici, une relecture a lieu
- A_pret_passe2.xml

Etapes de la fabrication - 3 Deuxième passe d'annotation

- Des éléments plus difficiles à repérer sont annotés.
- L'annotation est permise grâce à la mise en forme du texte original mais aussi grâce à la pose des balises de la passe précédente

Etapes de la fabrication - 4 Deuxième passe d'annotation

- Les éléments annotés sont les suivants :
  - L'étymologie : l'ensemble est groupé dans <etym>
    - <lang>
    - <gloss>
    - <mentioned>
  - Les renvois
    - <xr>
    - <lbl>
    - <ref>

Etapes de la fabrication - 5 Deuxième passe d'annotation

- Les éléments annotés sont les suivants :
  - L'étymologie
  - Les renvois
  - Proverbes et exemples
    - <cit type=« example »>
    - <cit type=« prov »>
    - <quote>
- Une relecture a lieu, dite « deuxième relecture couleur »
- A_pret_passe3.xml

Etapes de la fabrication - 6
Troisième passe d'annotation

- On annote les diverses sous-entrées :
  - <re type=« derive »>
  - <re type=« exp »>
  - <re type=« pronominal »>
- On annote (enfin!) les définitions (le plus difficile à repérer !)
  - Définitions classiques <def>
  - Définitions encyclopédiques <def type=« encycl. »>
- Une troisième relecture couleur a lieu
- A_definitif.xml

Les difficultés de l'annotation

- A première vue la forme du texte est régulière et correspond à son sens
- En réalité, ce n'est pas le cas
- Voici quelques exemples de détails :
  - Les abréviations annoncées ne sont pas celles utilisées
  - de temps en temps, les prononciations sont entre crochet et non plus entre parenthèses
- Parfois, les difficultés sont bien pires

Les difficultés de l'annotation
Le cas particulier de l'étymologie

- En théorie, elle se trouve après l'information grammaticale et on a :
- Etymologie = (nom d'une langue en abrégé + l'étymon en italique + virgule + glose)
- La preuve :
- **ABJECT** (*ab-jèkt'*), **E** adj. (lat. *abjectus*, jeté hors.)
- **ABJURATION** (*ra-si-on*) n. f. (lat. *abjuratio*, reniement.)
- **ABNÉGATION** (*si-on*) n. f. (lat. *abnegatio*, action de nier.)

Les difficultés de l'annotation
Le cas particulier de l'étymologie

- En pratique :
- **ABLÉGAT** (*ga*) n. m. (préf. *ab*, et lat. *legatus*, envoyé**.**)
- **ABOLIR** v. a. (lat. *abolere.*)
- **ABOI** n. m. (de *aboyer.*)
- **ABÎME** n. m. (du gr. *a* priv., et *bussos*, fond.)
- **ABÉE** (*bé*) n. f. (du vx fr. bée, auj. *baie*, ouverture.)
- **ABRÉGER** (*jé*) v. a. (lat. *abbreviare ;* de *brevis*, court. - Prend un e ouvert devant une syllabe muette :…

Les difficultés de l'annotation

Une solution : faire preuve d'imagination !

Pour certains cas, il faut adapter le script à plusieurs cas de figure :

- **ABÎME** n. m. (du gr. *a* priv., et *bussos*, fond.)
- **ABÉE** (*bé*) n. f. (du vx fr. bée, auj. *baie*, ouverture.)
- **ABOLIR** v. a. (lat. *abolere.*)
- **ABOI** n. m. (de *aboyer.*)
- **ABRÉGER** (*jé*) v. a. (lat. *abbreviare ;* de *brevis*, court. - Prend un e ouvert devant une syllabe…

Pour d'autre cas, il faut inventer : ici, nous avons typé la balise <etym> avec un type = « morpho » qui permet de distinguer une étymologie pure d'un découpage morphologique

- **ABLÉGAT** (*ga*) n. m. (préf. *ab*, et lat. *legatus*, envoyé.)

Les relectures

- Mise au point par Audrey d'un système astucieux de relecture double
- La relectrice relit le fichier grâce à une relecture « couleur ».
- Le fichier xml est affiché dans un navigateur web grâce à une feuille de style css, qui met en couleur les éléments à relire
- Parallèlement le fichier xml est relu dans un éditeur XML (type oxygen, notepad++) qui permet de relire directement les balises

Les relectures (2)

- Ce système, bien que très long, permet aux relectrices de se focaliser sur quelques éléments seulement et non sur la totalité du fichier.
- Il présente bien sûr l'inconvénient d'être long (on relit en alternance du texte et des balises)
- Il oblige à au moins quatre relectures par fichier (ce qui malheureusement n'est pas trop)
- A_pret_passe2.xml

Insertion des images

- Certaines images ont été scannées avec le texte et extraites automatiquement au moment de la conversion en HTML
- La plupart (9 sur 10) sont scannées manuellement et mises en ligne grâce à l'interface d'administration mise au point par Audrey
- http://bruscand.audrey.free.fr/petit-larousse-helene/admin

Présentation de la base

Allons visiter !

Conclusions

- La nouvelle base lexicale est sur le point d'exister
- L'utilisation de la TEI est possible même sur une base lexicale ancienne, conçue bien avant l'idée même de l'informatique (à condition d'inventer quelques types)
- Le balisage automatique est possible, à condition d'y ajouter une supervision humaine permanente à chaque étape

Perspectives

- A très court terme :
  - Ajout des lettres manquantes (texte et images) et mise en ligne officielle
  - Ajout d'un mode d'emploi et des publications liées à la ressource sur le site
- A moyen terme :
  - passe supplémentaire de corrections
  - Amélioration du moteur de recherche (qui ne fonctionne que sur des mots entiers pour le moment)