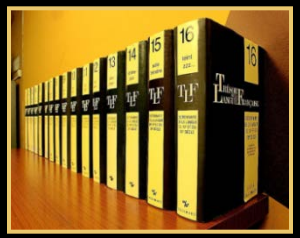
atilf
ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

Dictionnaires informatisés : les pratiques à l'ATILF

Environnement numérique, standardisation des langues et langue basque
Saint-Sébastien 12-13 septembre 2019

gilles.souvay@atilf.fr


www.atilf.fr

cnrs dépasser les frontières

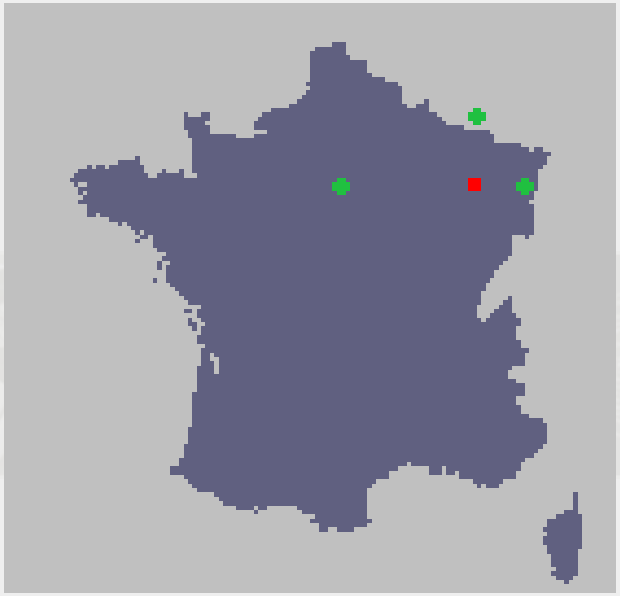
UNIVERSITÉ DE LORRAINE

Plan

- ATILF
- Ressources informatisées
- Dictionnaire du Moyen Français
- Conclusion

ATILF

- Analyse et Traitement Informatique de la Langue Française
  - Unité Mixte de Recherche
  - CNRS : Centre National de la Recherche Scientifique SHS
  - Université de Lorraine
  - www.atilf.fr



- Nancy

Plan

- ATILF
- Ressources informatisées
  - lexicographie (dictionnaires en ligne)
  - bases de données textuelles
  - outils pour le TAL
- Dictionnaire du Moyen Français
- Conclusion

Lexicographie informatique

- Dictionnaires électroniques en ligne
  - une des spécialités du laboratoire ATILF
    - français contemporain et langue médiévale
    - étymologie
  - références dans leur domaine
    - contenu scientifique
    - méthodologie

Lexicographie informatique

- Mode image // mode texte
  - plein texte // texte structuré
- Accès aux entrées
  - défilement des pages
  - gestion de la flexion – connaissance de la langue
  - prise en compte de la variation graphique
- Recherches dans le corps des articles
  - mot brut et ses flexions
  - structure de l'article

Lexicographie informatique

- Combiner des critères de recherche
- Connexion à d'autres ressources
- Dictionnaire figé // dictionnaire évolutif
  - autonomie informatique du responsable du projet
- Sources des exemples
  - création des exemples
  - corpus, bibliographie

Trésor de la Langue Française

- TLF : Trésor de la Langue Française
  - français du XIX^{e} et XX^{e} siècles
  - 100 000 mots avec leur histoire
  - 270 000 définitions
  - 430 000 exemples
  - 350 millions de caractères

- Version papier en 16 volumes
  - 1960->1990

Un article

PENSÉE[1], subst. fém. | **PENSÉE[2]**, subst. fém. | PENSER[1], verbe

A. – *BOT.* Plante herbacée, de la famille des Violacées, annuelle ou vivace, aux fleurs veloutées tricolores, comptant de nombreuses espèces sauvages ou cultivées, considérée comme le symbole du souvenir. *Pensée des champs; bourriche de pensées. Les alouettes sont partout ce soir. Il y a déjà des pensées bleues, blanches, violettes, par bandes dans l'herbe* (Pourrat, *Gaspard*, 1931, p.109):

- 1. Des **pensées**, des oeillets, des ravenelles, quelques rosiers, agonisaient au fond de ce puits sans air et chauffé comme un four par la réverbération des toits. Maupass., *Contes et nouv.*, t.1, Dimanches bourg. Paris, 1880, p.296.

B. – *P. méton.*

1. La fleur elle-même. *Sur le piédestal s'accumulent les humbles couronnes et les petits bouquets d'immortelles et de pensées* (Ménard, *Rêv. païen*, 1876, p.218). *Yeux en grand deuil violet comme des pensées!* (Laforgue, *Poés.*, 1887, p.207). V. *émail* ex. 3.

– [Dans son utilisation médicinale] *Le docteur s'est mis à rire de mes craintes (...) «Tant que vous mènerez votre chaste vie monacale et que vous travaillerez douze heures par jour, prenez tous les matins une infusion de* pensée sauvage*»* (Balzac, *Lettres Étr.*, 1834, p.129).

2. Représentation stylisée de cette fleur (relativement au symbole qu'elle évoque). *Un médaillon encadré qui contenait une pensée dessinée en cheveux rouges* (Champfl., *Avent. M^lle Mariette*, 1853, p.175). *Dans la paume qu'elle lui tendait il mit une pensée de saphir* (L. de Vilmorin, *Belles am.*, 1954, p.213). V. *médaillon* ex. 2.

3. *En empl. apposé inv.* [Désigne la nuance violet pourpre d'une variété de pensée] *Ruban pensée. Elle portait un tablier de soie violet pensée, avec la bavette* (Sand, *Mare au diable*, 1846, p.200). *Des velours ramagés, couleur pensée, formaient les rideaux et les portières* (Bourges, *Crépusc. dieux*, 1884, p.102):

- 2. *Toilette de visite (chapeau Figaro).* –Jupe en faille **pensée** avec un grand volant, haut sur la traîne et bas devant. Tunique en velours **pensée** garnie de plumes bleu très-pâle, et de brandebourgs en passementerie **pensée**. Manches ornées de crevés de satin **pensée**. Chapeau Figaro en velours **pensée** et bleu très-clair. Mallarmé, *Dern. mode*, 1874, p.779.

Prononc. et Orth.: [pɑ̃se]. Homon. *panser*. Att. ds *Ac.* dep. 1694. **Étymol. et Hist.** 1460-66 (Martial d'Auvergne, *Arrêts d'amours*, éd. J. Rychner, p.36, 65). De *pensée[1]**, cette fleur étant considérée comme l'emblème du souvenir, *cf.* 1558 *herbe de la pensée* (L. Fuchs, *Histoire des plantes* ds Roll. *Flore* t.2, p.173).

Trésor de la Langue Française informatisé

- TLFi
- Structuration fine des données
  - différents éléments de l'article
  - définition, conditions d'emploi, domaines, (exemples, sources)...
- On balise les informations
  - balise ouvrante <definition>
  - balise fermante </definition>

Un article

PENSÉE[1], subst. fém. | **PENSÉE[2]**, subst. fém. | PENSER[1], verbe

A. – *BOT.* Plante herbacée, de la famille des Violacées, annuelle ou vivace, aux fleurs veloutées tricolores, comptant de nombreuses espèces sauvages ou cultivées, considérée comme le symbole du souvenir. *Pensée des champs; bourriche de pensées. Les alouettes sont partout ce soir. Il y a déjà des pensées bleues, blanches, violettes, par bandes dans l'herbe* (POURRAT, *Gaspard*, 1931, p.109):

- 1. Des **pensées**, des oeillets, des ravenelles, quelques rosiers, agonisaient au fond de ce puits sans air et chauffé comme un four par la réverbération des toits. MAUPASS., *Contes et nouv.*, t.1, Dimanches bourg. Paris, 1880, p.296.

B. – *P. méton.*

1. La fleur elle-même. *Sur le piedestal s'accumulent les humbles couronnes et les petits bouquets d'immortelles et de pensées* (MÉNARD, *Rêv. païen*, 1876, p.218). *Yeux en grand deuil violet comme des pensées!* (LAFORGUE, *Poés.*, 1887 p.207). V. *émail* ex. 3.

– [Dans son utilisation médicinale] *Le docteur s'est mis à rire de mes craintes (...) «Tant que vous mènerez votre chaste vie monacale et que vous travaillerez douze heures par jour, prenez tous les matins une infusion de pensée sauvage»* (BALZAC, *Lettres Étr.*, 1834, p.129).

2. Représentation stylisée de cette fleur (relativement au symbole qu'elle évoque). *Un médaillon encadré qu contenait une pensée dessinée en cheveux rouges* (CHAMPFL., *Avent. M^lle Mariette*, 1853, p.175). *Dans la paume qu'elle lui tendait il mit une pensée de saphir* (L. DE VILMORIN, *Belles am.*, 1954, p.213). V. *médaillon* ex. 2.

3. *En empl. apposé inv.* [Désigne la nuance violet pourpre d'une variété de pensée] *Ruban pensée. Elle portait un tablier de soie violet pensée, avec la bavette* (SAND, *Mare au diable*, 1846, p.200). *Des velours ramagés, couleu pensée, formaient les rideaux et les portières* (BOURGES, *Crépusc. dieux*, 1884, p.102):

- 2. *Toilette de visite (chapeau Figaro).* –Jupe en faille **pensée** avec un grand volant, haut sur la traîne et bas devant. Tunique en velours **pensée** garnie de plumes bleu très-pâle, et de brandebourgs en passementerie **pensée**. Manches ornées de crevés de satin **pensée**. Chapeau Figaro en velours **pensée** et bleu très-clair

Prononc. et Orth.: [pɑ̃se]. Homon. *panser*. Att. ds *Ac.* dep. 1694. **Étymol. et Hist.** 1460-66 (MARTIAL D'AUVERGNE, *Arrêts d'amours*, éd. J. Rychner, p.36, 65). De *pensée*[1]*, cette fleur étant considérée comme l'emblème du souvenir, *cf.* 1558 *herbe de la pensée* (L. FUCHS, *Histoire des plantes* ds ROLL. *Flore* t.2, p.173).

TLF

PENSÉE[1], subst. fém. | **PENSÉE**[2], subst. fém. | PENSER[1], verbe

A. – *BOT.* Plante herbacée, de la famille des Violacées, annuelle ou vivace, aux fleurs veloutées tricolores, comptant de nombreuses espèces sauvages ou cultivées, considérée comme le symbole du souvenir. *Pensée des champs; bourriche de pensées. Les alouettes sont partout ce soir. Il y a déjà des pensées bleues, blanches, violettes, par bandes dans l'herbe* (Pourrat, *Gaspard*, 1931, p.109):

```
<Article>
<Entrée>PENSÉE</Entrée>, <CodeGrammatical>subst. fém.</CodeGrammatical>
<Paragraphe><Numéro>A. – </Numéro><Domaine>BOT. </Domaine><Définition>Plante herbacée, de la famille des Violacées, annuelle ou vivace, aux fleurs veloutées tricolores, comptant de nombreuses espèces sauvages ou cultivées, considérée comme le symbole du souvenir.</Définition><Syntagme>Pensée des champs</Syntagme>; <Syntagme>bourriche de pensées.</Syntagme> <Exemple>Les alouettes sont partout ce soir. Il y a déjà des pensées bleues, blanches, violettes, par bandes dans l'herbe <ReferenceBibliographique>(Pourrat, Gaspard, 1931, p.109);</ReferenceBibliographique></Exemple></Paragraphe>
...
</Article>
```

Trésor de la Langue Française

- Schéma de saisie des données
- Langages de balisage
  - XML : eXtensible Markup Language
  - définir son propre schéma
  - suivre des recommandations TEI (Text Encoding Initiative)
- Méthode de balisage
  - saisie initiale balisée avec un éditeur spécialisé
  - balisage manuel // balisage automatique

TLFi

- Versions informatisées
  - cédérom *version obsolète*
  - www.atilf.fr/tlfi
    - ergonomie des interfaces et technologie vieillissante
    - la plus élaborée dans les différents types de recherche
  - www.cnrtl.fr
    - recherche sur les entrées uniquement
    - connexion avec d'autres ressources

Linguistique historique française et romane

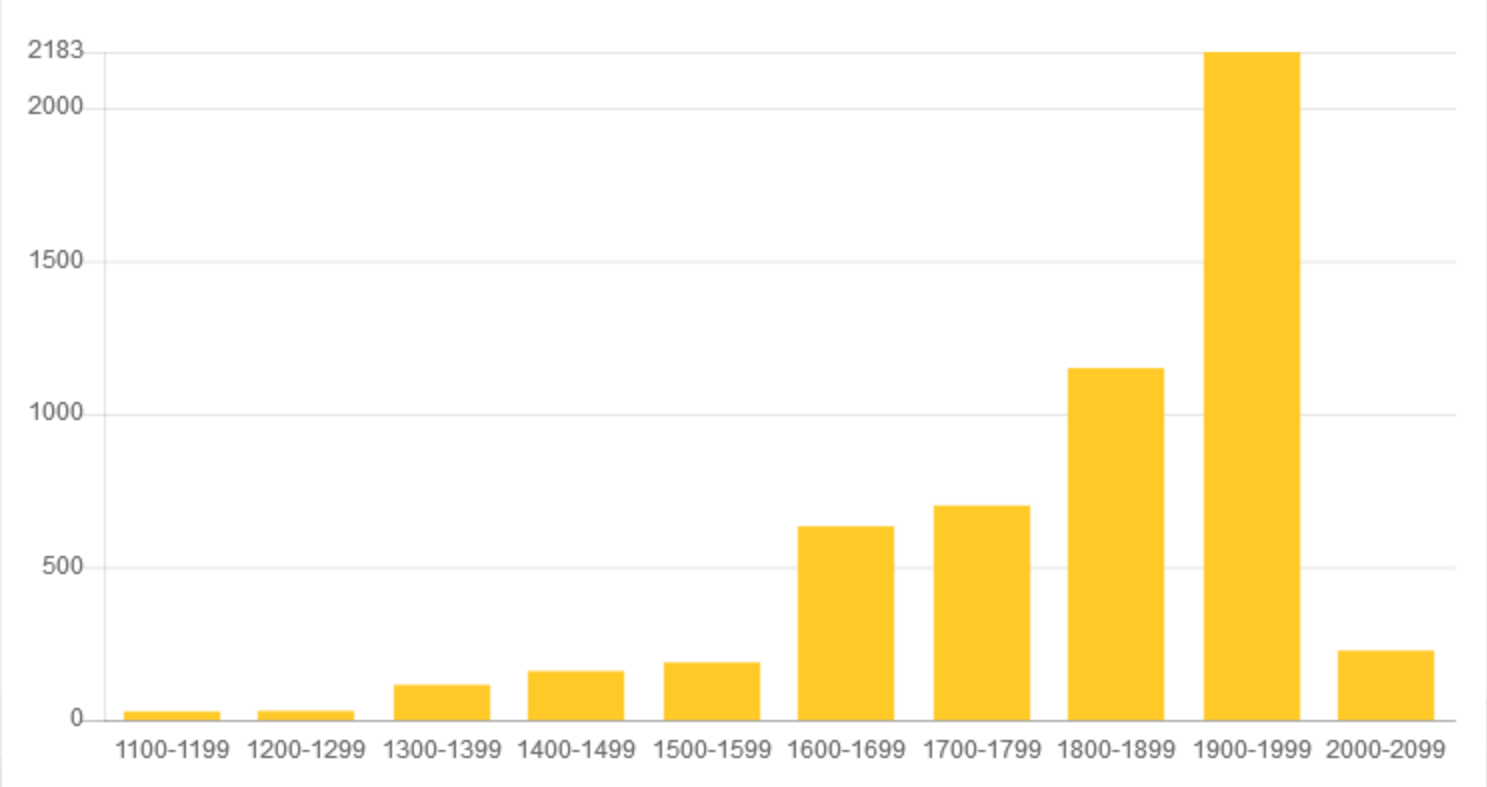
- Étymologie
  - FEW : Französisches Etymologisches Wörterbuch
  - TLF-Étym : révision sélective des notices étymologiques du TLFi
  - DÉRom : Dictionnaire Étymologique Roman

- Diachronie / Synchronie
  - Histoire de la langue et ses évolutions
  - Dictionnaire Électronique de Chrétien de Troyes (12^{e}s.)
  - Dictionnaire du Moyen Français (1330-1500)

Bases de données textuelles FRANTEXT

- https://www.frantext.fr
- Corpus créé dans les années 70 afin de fournir des exemples pour le *Trésor de la Langue Française*
  - textes littéraires et philosophiques
  - 5 415 références (juin 2019)
- Base de données de textes en français
  - différentes déclinaisons
  - accès libre ou sur abonnement

Frantext : qu'est-ce donc ?



Outils TAL

- Outils pour le Traitement Automatique de la langue
  - lexiques morphologiques
  - étiquetage de corpus, lemmatisation
  - ...

Plan

- ATILF
- Ressources informatisées
- Dictionnaire du Moyen Français
  - langue non normée
  - lemmatisation
- Conclusion

Dictionnaire et variation

- Comment trouver un mot dans le dictionnaire
  - dictionnaire langue contemporaine vs dictionnaire en diachronie
    - variation morphologique
    - variation graphique : norme / absence de norme
- Problème pour l'utilisateur
  - spécialiste ou pas de la langue médiévale
- Quelle entrée ? (Quel lemme ?)
  - *destroict*, *ameroyent*, *acoremens*, *polra*, *aulter*

Dictionnaire et variation

- Choix de la graphie du lemme
  - problème aussi pour le rédacteur
    - agnel, aigneau, agneau
    - DMF : moderniser
      - cohérence dans une famille, mots disparus
- Consultation du DMF
  - lemmatisation à la volée du mot
  - taper la forme telle que rencontrée dans le document, le DMF fera des propositions

Exemple de forme atypique

Formulaire

embache | Rechercher | Effacer | - options

☑ lemmatiser ☐ développer une graphie connue ☑ trace **LGeRM**
☐ attestation dans les corpus textuels
☑ analyse dans la *Base de Graphies Verbales*
☑ afficher les dictionnaires cités

Saisir un mot ou une forme sans se préoccuper des entrées du DMF : des propositions s'afficheront.

La recherche porte sur les variantes graphiques connues du lemmatiseur.

Résultat de la recherche

La forme *embache* est connue du lemmatiseur avec l'analyse suivante :

EMBATTRE, verbe | structure | sans exemple | complet | textes

[TL : *embatre* ; GD : ***embatre*** ; AND : ***enbatre1*** ; DÉCT : ***embatre*** ; FEW I, 293a ***battuere*** ; TLF : ***embat(t)re***]

Plus d'hypothèses

BGV

2 attestations dans la **Base de Graphies Verbales** http://www.atilf.fr/bgv/

embache	embatre	subjonctif présent 3	TL
embache	embatre	subjonctif présent 3	Gdf

Dictionnaire du Moyen Français (1330-1500)

- L'informatique est au cœur de l'élaboration du dictionnaire :
  - conservation et sélection des exemples dans des bases de données
  - rédaction des articles
- Logiciel de saisie
  - on n'utilise plus de traitement de texte
  - éditeur de texte balisé

Dictionnaire du Moyen Français (1330-1500)

- En accès libre : www.atilf.fr/dmf
- DMF 2015
  - 65 720 entrées, 470 125 exemples
  - 200 millions de caractères
  - 19 900 pages, 15 volumes du TLF
- Points forts
  - bonne couverture de la langue médiévale
  - bonne gestion de la variation graphique
  - un dictionnaire ‘hyperconnecté’
  - un dictionnaire ‘hyperconnectable’

Dictionnaire du Moyen Français (1330-1500)

FIEF, subst. masc. **fief**
[T-L, GDC : ***fief*** ; FEW XV-2, 117a : ****fehu*** ; TLF VIII, 843b : ***fief***]
A. - DR. FÉOD.
1. "Domaine noble relevant d'un suzerain que celui-ci concède en tenure à un vassal (en dehors de toute rente) en contrepartie de l'hommage et du service requis" : Del ***fief*** l'empereor estes a tort saisis (Garin Lorr. M., c.1330-1400, 486). ...deux cenz livres de rente, les quelles le dit conte avoit vendues au dit cardinal, assises, selon la coustume du païs, en la chastelerie de Syvray, avec toute justice, haute, moienne et basse en ***fiez*** et arrerefiez, et à touz autres droiz, quiex qu'il feussent, excepté tant seulement ressort et souveraineté (Doc. Poitou G., t.2, 1335, 125). ...je suis si courcié que sommes si meschant Que n'ay terre, ***fief***, ne ung chastel vaillant (Ren. Gennes D.B., c.1350-1400, 89). Vo ***fief*** en croisteray d'une riche contree. (Renaut Mont. B.N. V., c.1350-1400, 366). Tant avoit richesse et puissance, Terres, ***fiez*** [var. ***fies***], honneur et avoir Que trop estoit de tant avoir. (MACH., C. ami, 1357, 29). Il donnoit ***fiez***, joiaus et terre, Or, argent ; riens ne retenoit Fors l'onneur ; ad ce se tenoit, Et il en avoit plus que nuls. Des bons fu li mieudres tenus. (MACH., C. ami, 1357, 103). ...mon nepveu (...) Met si a non chaloir le sien Que de ses ***fiez*** et heritages Ne li chaut (Mir. chan., c.1361, 141). Et certes, selon la Loy civile, et selon raison, le vassal qui ne fait le service que il doit a

Dictionnaire du Moyen Français (1330-1500)

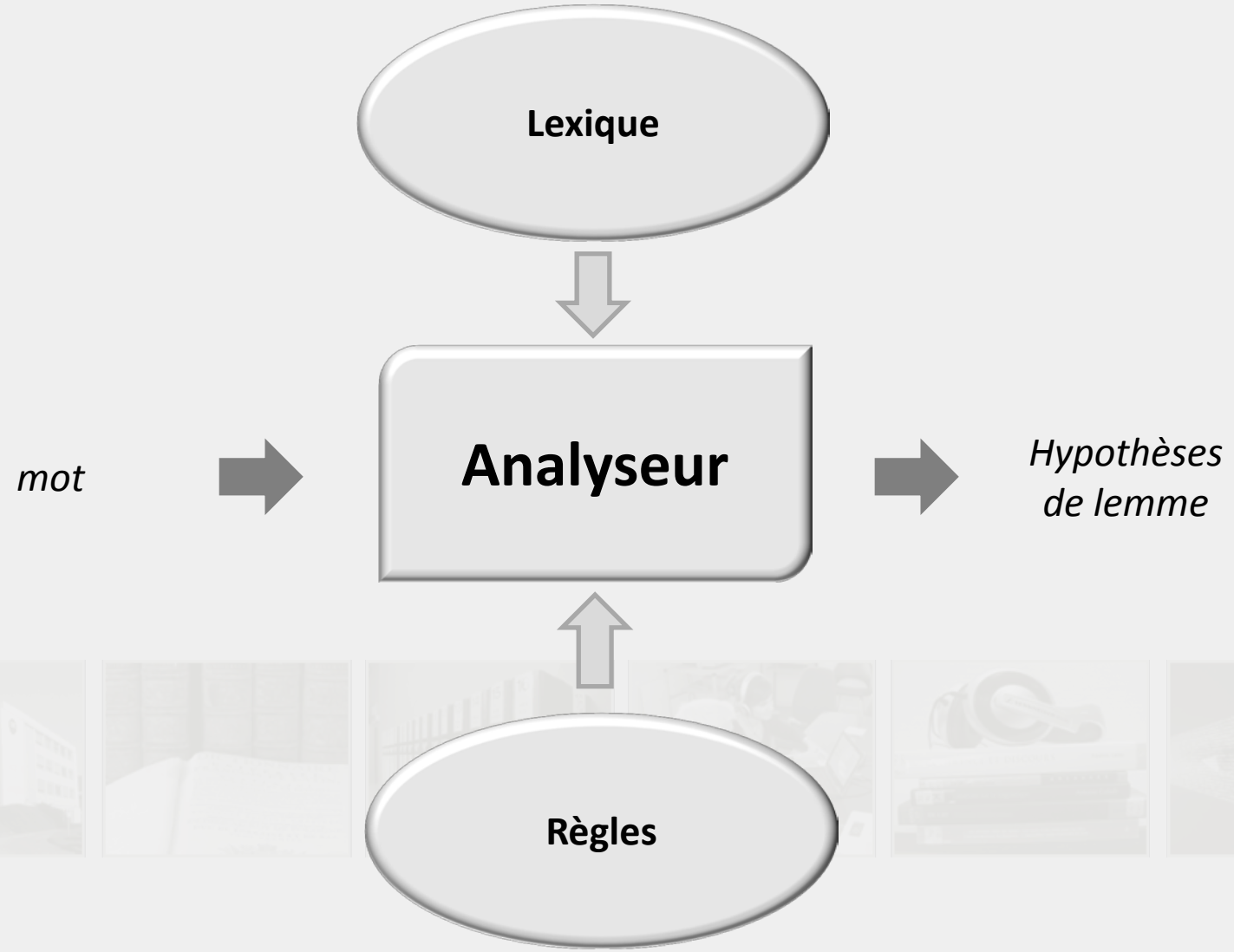
FIEF // *FEUDO*

ART VED **FIEF** /VED CODE , **subst. masc.** /CODE LEM **fief** /LEM
DICT [TLGDC T-L, GDC : LEMME ***fief*** /LEMME /TLGDC FEW.CONNU ; FEW VOLUME XV-2, /VOLUME PAGE 117a : /PAGE ETYM ****fehu*** /ETYM /FEW.CONNU TLF ; TLF VOLUME VIII, /VOLUME PAGE 843b : /PAGE LEMME ***fief*** /LEMME /TLF] /DICT
P DISC NUM A. - /NUM DOM DR. FÉOD. /DOM /DISC /P
P DISC NUM **1.** /NUM DEF "Domaine noble relevant d'un suzerain que celui-ci concède en tenure à un vassal (en dehors de toute rente) en contrepartie de l'hommage et du service requis" /DEF /DISC EXE :
TEXTE Del OCC ***fief*** /OCC l'empereor estes a tort saisis /TEXTE REF (Garin Lorr. M., c.1330-1400, 486) /REF. /EXE EXE TEXTE ...deux cenz livres de rente, les quelles le dit conte avoit vendues au dit cardinal, assises, selon la coustume du païs, en la chastelerie de Syvray, avec toute justice, haute, moienne et basse en OCC ***fiez*** /OCC et arrerefiez, et à touz autres droiz, quiex qu'il feussent, excepté tant seulement ressort et souveraineté /TEXTE REF (Doc. Poitou G., t.2, 1335, 125) /REF. /EXE EXE TEXTE ...je suis si courcié que sommes si meschant Que n'ay terre, OCC ***fief*** /OCC, ne ung chastel vaillant /TEXTE REF (Ren. Gennes D.B., c.1350-1400, 89) /REF. /EXE EXE TEXTE Vo OCC ***fief*** /OCC en croisteray d'une riche

Dictionnaire du Moyen Français

- LGeRM
  - **L**emmes, **G**raphies **e**t **R**ègles **M**orphologiques
  - Point d’entrée dans le DMF

Architecture



Analyseur

Algorithme

```
si la graphie est dans le lexique alors
    Proposer l'analyse
sinon
  tantque conditions faire
      Appliquer les règles pour trouver une graphie connue
  fait
finsi
```

Conditions

- mécanisme d'arrêt
- stratégie de gestion des formes produites

Le lexique

- Liste de triplets (forme, lemme, étiquette)
  - (*amer*, AIMER, verbe)
  - (*amer*, AMER, adj.)
  - (*amera*, AIMER, verbe)

- Construction du lexique
  - lexique initial issu des exemples du DMF, lemmes et étiquettes
  - enrichi à partir du corpus FRANTEXT
  - de collaborations
  - août 2019 environ 975 500 entrées

Les règles 1/5

- Règle (morphologique) : morphologie et variation graphique
- Structure générale d'une règle
  - `Si` *conditions* `alors` *action* `finsi`
- Conditions sur les graphèmes du mot
  - en finale, en initiale
  - précédé de, suivi de : une lettre, liste de lettres, d'une consonne, d'une voyelle, sauf ...
- Conditions sur le lemme
- Conditions sur le succès de la règle

Les règles 2/5

- Règles sur la flexion verbale
  - finale : retrouver l'infinitif
  - transformation de la finale : autre personne
    - si (en finale) alors RONT → RA finsi
    - *menront* → *menra*, MENER
  - autre transformation de la finale
    - si (en finale) et (précédé de [D,T,V]) alors ERAI → RAI finsi
    - *ponderai* → *pondrai*, PONDRE
  - ou cas inverse
    - *menront* → *meneront*

Les règles 3/5

- Flexion non verbale
  - si (en finale) alors ES → EFS finsi — *nes*→ *nefs*, NEF
- Modernisation/archaïsation
  - Y → I — *fayre*→ *faire*, FAIRE
- Équivalence graphique
  - C→ SS — *mesfacent* → *mesfassent*, MÉFAIRE
- Agglutination adverbe, pronom, élément formant
  - *tresadvisé* → *advisé*, TRÈS +AVISÉ
- Variantes régionales
  - OUN → ON — *mount*→ *mont*, MONT

Les règles 4/5

- Le système comporte environ 6 500 règles
  - 200 règles initiales de 1986
  - ajout de la flexion verbale en 2001
  - confrontation au DMF et aux corpus textuels
  - ¾ pour la flexion verbale et ses variations
- Pallier les lacunes du lexique
  - On n'aura jamais toutes les variations possibles d'un mot dans le lexique
  - la variation graphique est contenue dans les règles

Les règles 5/5

- Décrire la variation/flexion du lemme CONNAISSANCE

[c|k|q][o|oi|e|oei][n|nn|gn|ngn][oi|ai|i|ioi|e|oe][s|ss|sc|sç|ç|c][i]?[en|an|ã|ẽ][s|ss|c|sc|ç|ch][e][sz]?

cognescence cognissance cognissanche
cognoeissance cognoiscences cognoisçance conaisanche
congnioissance congnoessance
connissanche conoissances cougnoissance...

- 55 formes attestées dans nos corpus médiévaux

Prise en compte

- Toute variante graphique à partir du moment où le lemme est connu
  - enrichissement permanent
    - liste de formes
    - liste de règles
    - liste des lemmes
- Capable de traiter une transcription diplomatique
  - u/v i/j barre de nasalisation s long
  - *ſciẽce, neceßité, comẽ*

Produits dérivés

- Lexiques de formes
  - LGeRM médiéval (juin 2016)
    - optimisé pour 1300-1500
    - 88 039 lemmes et étiquettes DMF
    - 951 452 entrées / 192 607 attestées FRANTEXT
  - LGeRM XVI-XVIIe (2013) / LGeRM PRESTO
    - optimisé pour 1550-1700
    - 89 754 lemmes et étiquettes TLF
    - 2 959 371 entrées / 116 161 attestées (3,9%)
- Lexiques de modernisation

- Dictionnaire entièrement informatisé
  - structure fine en XML
  - outils d'aide à la rédaction
    - contrôle des articles en ligne
    - bibliographie
  - administration du projet
    - dépendant de l'informaticien
    - DÉROM : la responsable du projet gère les mises à jour

Plan

- ATILF
- Ressources informatisées
- Dictionnaire du Moyen Français
- Conclusion

Conclusions

- Expérience de l'ATILF
  - dictionnaires informatisés
  - lemmatisation
- Modèle pour d'autre projets
- Outils de références
  - TLFi, DMF, DÉRom...
  - corpus textuel Frantext