La polysémie évolutive du lexique français (12e -20e siècle)

Laurette Chardon, Justine Reynaud, Jacques François

Cahier du CRISCO
n°37
septembre 2023

LA POLYSÉMIE ÉVOLUTIVE DU LEXIQUE FRANÇAIS

(12e–20e siècle)

Projet d'informatisation fonctionnelle et de modélisation graphique des entrées historiques du TLFi

Laurette CHARDON
CRISCO

Justine REYNAUD
GREYC

Jacques FRANÇOIS
CRISCO

Université de Caen (Bât. Sciences Porte SA S13), 14032 CAEN CEDEX
Tél. : 02 31 56 56 27 — Fax : 02 31 56 54 27 — Site web : https://crisco.unicaen.fr
Courriel direction : thierry.ruchot@unicaen.fr
Courriel secrétariat : nelly.donnet@unicaen.fr

Résumé

Cette étude est consacrée à l'exploration de la notion de « polysémie évolutive », qui ne pointe pas sur un type particulier de polysémie lexicale, mais sur l'évolution de l'éventail des sens véhiculés par un vocable au fil des siècles, essentiellement par le biais d'extensions de sens, car le déclin d'une partie d'entre eux se laisse difficilement documenter. L'accrétion progressive des sens des vocables polysémiques en français contemporain a fait l'objet d'études approfondies depuis plus d'un siècle, depuis les travaux d'A. Darmesteter (1887) jusqu'au dictionnaire historique d'A. Rey (1992) en passant par le FEW de W. von Wartburg (depuis 1929) et les entrées « Étymologie et Histoire » du Trésor de la Langue Française rédigées dans les trois dernières décennies du 20^{e} siècle sur cette base. L'informatisation du TLF a donné lieu à deux niveaux de traitement : formel et fonctionnel (les différentes polices de caractères et leur fonction contextuelle) pour les entrées principales, mais seulement formel pour les entrées historico-étymologiques, ce qui interdit toute enquête transversale. L'objet de notre étude, assise sur trois publications antérieures (François 2020a, 2020b, 2021) est double. Il s'agit d'abord de compléter le traitement formel de ces entrées en attribuant une fonction lexicographique diachronique à différents types de segments, notamment les identifiants des multiples rubriques (jusqu'à 154 pour le verbe *tenir*), la datation de la première attestation de chaque type d'emploi, les collocations en cause, les définitions, etc.). Ensuite ces classes de données sont rassemblées dans un tableau de conversion graphique, lequel permet de construire automatiquement un graphe historique arborescent pour les ±29 000 entrées polysémiques (sur ±49 000 au total). Le but de cette modélisation graphique est de permettre la comparaison entre des types de profils historiques dont la dernière section de l'article donne une illustration en combinaison avec l'exploitation diachronique de FRANTEXT du 12^{e} siècle à nos jours.

Mots-Clés — TLFi ; polysémie évolutive ; Étymologie et histoire ; traitement informatique XML ; modélisation graphique

Abstract

This study is devoted to exploring the notion of "evolving polysemy", which does not focus on a particular type of lexical polysemy, but on the evolution of the range of meanings conveyed by a word over the centuries, essentially through meaning extensions, as the decline of some of them is difficult to document. The gradual accretion of lexical polysemies in contemporary French has been the subject of in-depth study for over a century, from the work of A. Darmesteter (1887) to A. Rey's historical dictionary (1992), via W. von Wartburg's FEW (since 1929) and the "Etymology and History" entries in the Trésor de la Langue Française (TLF), which were written on this foundation in the last three decades of the 20th century. The computerization of the TLF has resulted in two levels of processing : formal and functional (the different fonts and their contextual function) for the main entries, but only formal for the historical and etymological entries, which precludes any cross-disciplinary investigation. The aim of our study, based on three previous publications (François 2020a, 2020b, 2021), is twofold. Firstly, the formal processing of these entries is completed by assigning a diachronic lexicographic function to various types of segments, including the identifiers of the multiple headings (up to 146 for the verb *tenir*), the date of the first attestation of each type of use, the collocations involved, the definitions, etc.). These data classes are then assembled in a graphical conversion table, which automatically constructs a tree-like historical graph for the ±29,000 polysemous entries (out of ±49,000 in total). The aim of this graphical modeling is to ease the comparison between different types of historical profiles, as illustrated in the last section of the article in combination with the diachronic mining of FRANTEXT from the 12th century to the present day.

Keywords — TLFi ; polysemy evolution ; etymology and history ; XML ; visual modeling

Table des matières

1 La phase préliminaire du projet (2020–21) : deux questionnements en voie de convergence 5
1.1 Techniques de visualisation de la diversification des sens lexicaux 5
1.2 Deux techniques en concurrence 10

2 Le projet d'informatisation fonctionnelle et de modélisation graphique des entrées historico-étymologiques [H-É] du TLFi 12

3 La microstructure des articles du TLFi et de leur entrée H-É 15
3.1 Articles, entrées, dégroupements et regroupements 15
3.2 Comparaison entre les deux entrées lexicographiques du H-É du n. m. CŒUR 16

4 L'éventail typographique des entrées H-É 19
4.1 La fonction variable des types de caractères selon qu'ils figurent entre ou en dehors de parenthèses courbes 20
4.2 La dissociation des deux sous-entrées historique et étymologique 22

5 Le traitement informatique des entrées H-É au format XML 23
5.1 Structure générale des fichiers 23
5.1.1 Les principales balises XML 23
5.2 La balise `<etymology>` 25
5.3 L'algorithme de création des données tabulaires 25
5.3.1 L'algorithme général 25
5.3.2 L'extraction des niveaux et des rubriques (proc 2) 26
5.3.3 L'extraction des niveaux et des rubriques (proc 3) 27
5.3.4 Identification de la définition et des collocations d'une rubrique (proc 4) 28
5.3.5 Séparation des parties Histoire et Étymologie (proc 5) 28

6 Bilan d'étape et illustration d'un champ d'étude dérivé 29
6.1 La poursuite du chantier en cours 29
6.2 Test de l'hypothèse d'une corrélation entre le volume de l'entrée H-É d'un vocable en nombre de rubriques et sa fréquence relative dans FRANTEXT au fil des siècles 30

Introduction

L'expression *polysémie évolutive* n'est pas destinée à pointer sur un type particulier de polysémie lexicale. *Évolutif* assume ici la fonction d'un « adjectif de nature », comme dans *Achille au pied léger* ou *la perfide Albion*, car toute polysémie lexicale résulte d'un jeu complexe d'accrétions sémantiques – ce qu'A. François (2008) appelle la 'colexification', néologisme qui a été repris en anglais par plusieurs autres linguistes depuis lors, et dont il explore la dimension diachronique sous le nom de « tectonique lexicale » (cf. A. François 2022). Il s'agira ici de porter notre attention sur la dimension évolutive de la polysémie des vocables [1] français, c'est-à-dire sur ses fluctuations, entre extensions de sens (cf. J. François, à par.), déclin, voire extinction de l'usage affectant des sens attestés par le passé, et périodes de stabilisation de ce que l'école allemande appelle le champ sémasiologique du vocable par opposition au champ onomasiologique des concepts qu'il est amené à véhiculer (cf. Blank 2000 ; Koch 2000).

Cette étude se situe dans le prolongement de trois articles préliminaires (J. François 2020a, 2020b, 2021) destinés à examiner la microstructure des articles du Trésor de la Langue Française informatisé [2] et plus spécifiquement celle de leurs entrées historico-étymologiques (désormais abrégé en H-É). Alors que les premières ont fait l'objet d'une informatisation minutieuse en langage SGML de la part de l'équipe de Jacques Dendien à l'INaLF dans la dernière décennie du 20e siècle (cf. Bernard, Dendien & Pierrel 2004 et Martin 2001), les secondes ont été laissées en friches. L'informatisation des premières est formelle (exploitant quatre types de caractères : romains, italiques, gras et petites capitales) et fonctionnelle (attribuant à chaque segment une fonction textuelle, ce qui permet de pratiquer des recherches transversales approfondies), mais celle des secondes n'est que formelle, ce qui exclut toute recherche transversale et donc toute exploitation avancée et toute modélisation graphique des entrées H-É.

L'objet de la recherche menée au CRISCO [3] depuis 2022 en collaboration entre un linguiste et deux informaticiennes, tous trois de l'université de Caen-Normandie, et avec le soutien de l'ATILF [4], est d'évaluer la nature et l'étendue des obstacles qui, il y a un quart de siècle, ont dissuadé les informaticiens de l'INaLF d'appliquer aux entrées H-É une procédure d'informatisation fonctionnelle comme dans les entrées principales, dites 'lexicographiques'. Ces entrées historiques ayant été rédigées au fil de trois décennies par différents rédacteurs qui ont manifestement reçu des consignes variables au fil du temps (notamment pour le niveau des identifiants des rubriques, parfois très nombreuses et très hiérarchisées, le degré de raffinement des références philologiques et la prise en compte des attestations isolées), l'impression première en comparant leur microstructure est effectivement décourageante, mais on s'aperçoit avec un peu plus d'entraînement, que le gabarit

1. Nous adoptons ici la distinction terminologique d'I. Mel'čuk et Alain Polguère entre *vocable* (cf. Mel' čuk 2023 : 274) « ensemble d'unités lexicales (= lexèmes ou idiotismes) dont les signifiants sont identiques et les signifiés partagent des composantes sémantiques importantes (= passerelles sémantiques) » et *lexème* (cf. Mel' čuk 2023 : 255), « L'un des deux types d'unité lexicale (l'autre étant les idiotismes) – une expression monolexicale »). Dans cette terminologie, chaque entrée H-É du TLFi a pour objet un vocable et chacune de ses rubriques un lexème, une collocation (cf. Mel' čuk 2023 : 248 « phrasème sémantico-lexémique compositionnel ») ou un idiotisme.

2. TLFi en ligne : https://www.cnrtl.fr/portailindex/LEXI/TLFI/A/80 ou https://www.cnrtl.fr/etymologie/vocable

3. *Centre de Recherches Interlangues sur la Signification en Contexte*, Université de Caen-Normandie.

4. Le laboratoire *Analyse et Traitement Informatique de la Langue Française* (CNRS et Université de Lorraine), qui a pris la suite de l'INaLF (*Institut National de la Langue Française*) au tournant du XXIe siècle.

de base de toutes les entrées H-É est globalement immuable. L'impression de variation radicale déjouant a priori toute tentative d'informatisation fonctionnelle vient en fait de la nécessité habituelle en lexicographie traditionnelle de gagner de l'espace et donc de pourchasser toute répétition dont on peut se dispenser. Par exemple, si sur un gabarit d'une quinzaine de classes de données bien ordonnées l'entrée H-É d'un vocable V1 mentionne les données des types A, B, E, G et I, tandis que celle d'un vocable V2 mentionne les données des types B, C, D, F, I et J, les données effectivement comparables se limitent à B et I au milieu de sept autres données incomparables, ce qui peut faire l'effet d'un chaos descriptif.

En fait la proportion des données comparables est généralement supérieure, mais à titre d'exemple il a été explicitement recommandé [5] aux rédacteurs, entre deux ou plusieurs rubriques successives d'un même niveau hiérarchique, de ne fournir qu'une seule définition commune (en caractères romains entre guillemets) et éventuellement d'ajouter des métadonnées [6] de domaine (ex. « mar. » pour le vocabulaire de la marine ou « fin. » pour celui des finances) ou de relation de sens (ex. « p.anal. » pour une relation d'analogie). De ce fait la recherche transversale de tous les segments entre guillemets fournit bien la liste de toutes les rubriques dotées d'une définition, mais d'une part certaines 'quasi-définitions' (du type : *désigne* ...) manquent à l'appel, ainsi que toutes les rubriques où la définition fait défaut et doit être recherchée dans la première rubrique de l'entrée ou de la sous-entrée en cause. Ce sont les types d'obstacles à l'informatisation fonctionnelle et donc à la modalisation graphique des entrées H-É que nous n'avons cessé de rencontrer mais qui, à force d'enregistrements cumulés ont fait finalement apparaître un ordre à peine perceptible à première vue.

Nous évoquerons en premier (§1) la phase préliminaire du projet (2020-21), puis la mise en œuvre du projet d'informatisation fonctionnelle et de modalisation graphique des entrées H-É, une fois que le trio d'investigation de ces entrées s'est mis en place entre l'hiver et le printemps 2022 (§2), l'analyse de la microstructure des ±20 000 entrées H-É décrivant une polysémie évolutive (§3), l'éventail typographique de ces entrées et leur interprétation fonctionnelle (§4), un aperçu des techniques de l'informatisation fonctionnelle des entrées H-É (§5) et un bilan d'étape complété par la présentation de matériaux pour l'examen de la corrélation présumée entre la polysémie évolutive d'un vocable selon le TLFi et ses occurrences de siècle en siècle dans FRANTEXT (§6).

1 La phase préliminaire du projet (2020–21) : deux questionnements en voie de convergence

Depuis une quinzaine d'années, deux axes de réflexion, indépendants l'un de l'autre à l'origine, se sont développés dans les travaux de Jacques François.

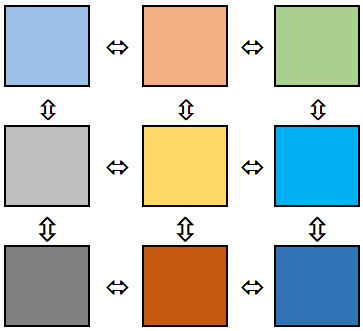
1.1 Techniques de visualisation de la diversification des sens lexicaux

Le premier axe porte sur la représentation visuelle des extensions de sens qui ont débouché sur la polysémie plus ou moins étendue de vocables français et notamment de

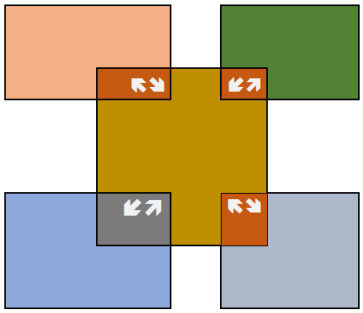
5. Nous remercions Eva Buchi d'avoir recherché à notre demande des instructions de rédaction des entrées H-É dans la bibliothèque de l'INaLF intégrée entre-temps dans celle de l'ATILF. Les instructions internes manuscrites qui ont été mises à jour datent de 1970-71 et mentionnent ces recommandations.

6. Cf. Caron, Defiolle et Lay (dir. 2019).

verbes particulièrement sujets à ce processus d'accrétion de sens successifs. Ces extensions s'expliquent par l'exigence d'économie lexicale dans la mémorisation des moyens d'expression des percepts qui, une fois structurés en concepts s'organisent en un réseau d'unités qui ne se distinguent pas radicalement (*versus* fig. 1a), mais qui présentent un degré variable de superposition (cf. fig. 1b).



(a) La face signifiée du lexique d'une langue vu comme un réseau de représentations sémantiques catégoriquement distinctes



(b) La face signifiée du lexique d'une langue vu comme un réseau de représentations sémantiques floues, c.à.d. comportant des recouvrements

FIGURE 1 – Deux visions de la face signifiée du lexique d'une langue.

Le premier sémanticien à avoir proposé une représentation visuelle de la superposition entre la couverture sémantique de différents vocables est sans doute Benjamin Lafaye dans la Préface à son Dictionnaire des synonymes (1858 : xxxix), cf. figure 2 ci-dessous.

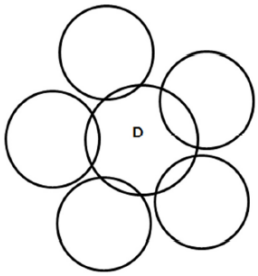


Figure 7 de Lafaye (1858) : idées communes et traits particuliers d'une famille de vocables de sens voisins. Le cercle central D, entrant en intersection avec les cinq autres, représente le sens commun aux cinq vocables, ce qui assure leur ressemblance. Si ce sens est véhiculé par le n.m. *malheur*, comme Lafaye le suggère, les vocables excentrés pourraient être par ex. *misère*, *infortune*, *échec*, *déconvenue*, *calamité*.

FIGURE 2 – Une visualisation sémantique de Benjamin Lafaye et son interprétation

Ce mode de visualisation permet effectivement de comprendre ce qui est en cause dans deux réalités linguistiques interdépendantes : la polysémie lexicale et la synonymie partielle. Ces deux réalités naissent d'une troisième, le déploiement des sens lexicaux dans des cotextes discursifs et des contextes énonciatifs. Au niveau des représentations mentales (celui de la CONCEPTUALISATION selon Levelt, 1989) les « traits sémantiques primitifs » (ang. *semantic primes*, cf. Wierzbicka 1996, etc.) ne peuvent pas avoir une silhouette parfaitement délimitée, ils ne sont accessibles que comme des types d'emploi valides d'au moins un vocable d'une langue particulière et généralement de plusieurs vocables qui partagent ce trait et sont ainsi substituables les uns aux autres dans un type de contexte particulier. Une variation plus sophistiquée de cette technique de visualisation figure dans l'article de J. François «Quand JOUER, c'est jouer de la musique» (2005 : 149, cf. figure 3 ci-dessous).

Dans la figure produite automatiquement par le DICTIONNAIRE ÉLECTRONIQUE DES SYNONYMES (dans sa version de l'époque), le verbe *jouer* occupe la même place centrale que le cercle D dans la figure 3 de Lafaye. Son « enveloppe convexe » est entourée de

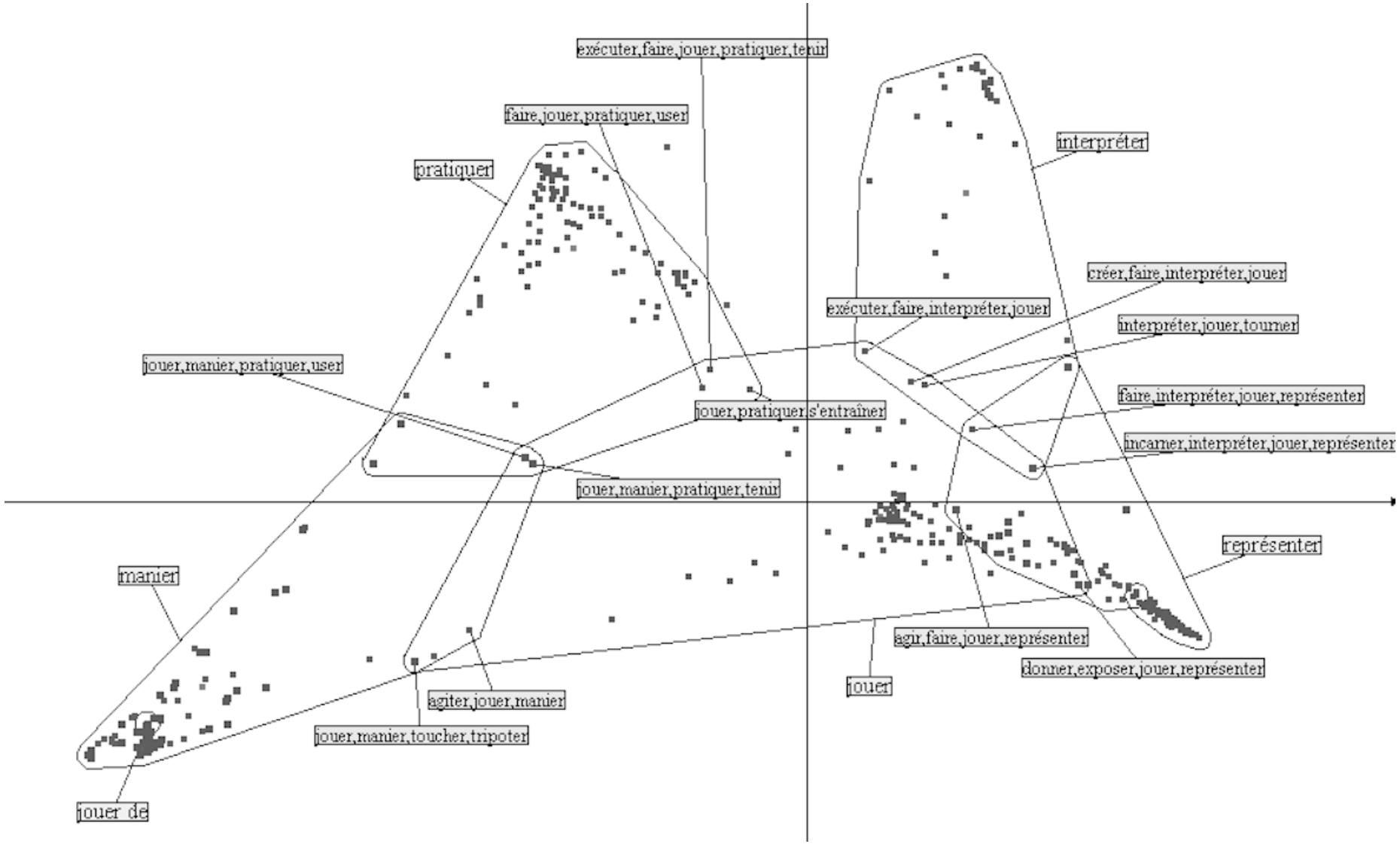


FIGURE 3 – Visualisation de l'intersection entre la polysémie du v. *jouer* et de 4 de ses synonymes contextuels {*interpréter, manier, pratiquer, représenter*} à l'aide du DICTIONNAIRE ÉLECTRONIQUE DES SYNONYMES du CRISCO.

celles de quatre de ses principaux synonymes (de gauche à droite : *manier, pratiquer, interpréter* et *représenter*). Le dictionnaire électronique dispose ces synonymes en fonction des synonymes que chacun des quatre partage avec le v. *jouer*. Les ensembles de synonymes partagés sont appelés des « cliques de synonymes ». Ainsi on voit que *jouer* et *manier* partagent 5 nouveaux synonymes qui figurent dans quatre cliques :

1. *agiter, jouer, manier*
2. *jouer, manier, pratiquer, tenir*
3. *jouer, manier, pratiquer, user*
4. *jouer, manier, toucher, tripoter*

Parmi ces cliques, la 2 et la 3 partagent également le 2ème synonyme, *pratiquer*. De même les deux cliques 5-6 sont partagées par le 3ème synonyme, *interpréter*, et le 4ème, *représenter*.

5. *faire, interpréter, jouer, représenter*
6. *incarner, interpréter, jouer, représenter*

En revanche les deux synonymes figurant à gauche dans la fig. 3 et les deux de droite n'ont aucun synonyme commun. La raison est facile à imaginer : en tant que synonymes de *jouer, manier* et *pratiquer* se combinent avec un complément d'objet désignant prioritairement un instrument de musique, tandis que *représenter* et *interpréter* prennent pour objet une œuvre théâtrale ou musicale ou un rôle dans une œuvre théâtrale ou lyrique. Le cotexte syntaxique est donc similaire pour *manier* ∼ *pratiquer* et pour *interpréter* ∼ *représenter*. La fig. 4 est une version simplifiée de la fig. 3 destinée à illustrer le contexte sémantaxique associé aux vocables membres de chacun des deux ensembles et l'appartenance de *jouer* aux deux ensembles.

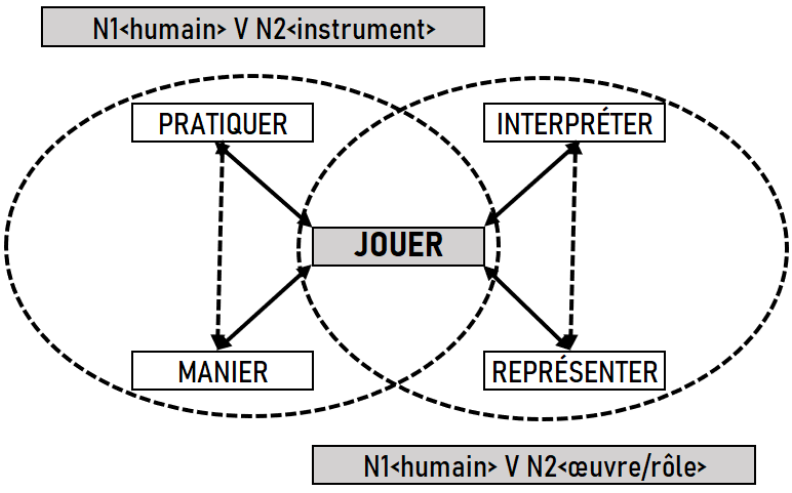


FIGURE 4 – Les deux ensembles de synonymes [*jouer, manier, pratiquer*] et [*jouer, interpréter, représenter*]

Passons maintenant à l'APPROCHE DIACHRONIQUE de la sémantique lexicale (cf. Blank 2000). La figure 5 reproduit un jeu de visualisations de trois types de polysémie évolutive emprunté à François (2008). Soit une unité lexicale L véhiculant à t^{α} un seul type de sens `<s1>`. À partir de t^{d} (époque de diversification sémantique) L peut véhiculer, selon le contexte, un second sens `<s2>`. À l'époque symbolisée par une droite rouge discontinue, les trois visualisations sont identiques. Au-delà, on est en présence de trois évolutions différentes relativement au moment d'observation t^{0}.

— Dans le premier cas, les deux types d'emploi sont toujours valides, c'est l'illustration d'une polysémie effective à t^0. L'ordre dans lequel les deux sens `<s1, s2>` sont apparus est généralement deviné par les usagers de la langue quand un facteur 'classique' de diversification des sens est intervenu (analogie de forme, de fonction, de constitution, etc., métaphore, métonymie, généralisation, spécialisation, cf. Paul 1880 ; Stern 1931 ; Ullmann 1951, 1962). Dans le cas inverse, cet ordre d'apparition leur reste inaccessible et ils jugent fréquemment qu'ils sont en présence de deux vocables homonymes.

— Dans le second cas, le premier type d'emploi a décliné et finalement disparu – sauf éventuellement sous la plume d'écrivains au style archaïsant – et l'usager de la langue à l'époque t^0 ignore généralement que L avait originellement un sens différent et est passé par une phase de polysémie avant que `<s2>` se substitue finalement à `<s1>`. Ainsi, qui sait encore que durant l'époque de la féodalité un seigneur **avouait** son suzerain, c'est-à-dire se reconnaissait comme son vassal ?

— Dans le troisième cas enfin, le sens `<s2>` a décliné et a disparu, tandis que `<s1>` s'est maintenu jusqu'à t^0. Il s'agit d'un cas de polysémie transitoire. La plupart des usagers de la langue ignore généralement que L a véhiculé également le sens `<s2>` pendant une période plus ou moins étendue (voir plus bas le sens 6 du n.m. *timbre* selon Darmesteter).

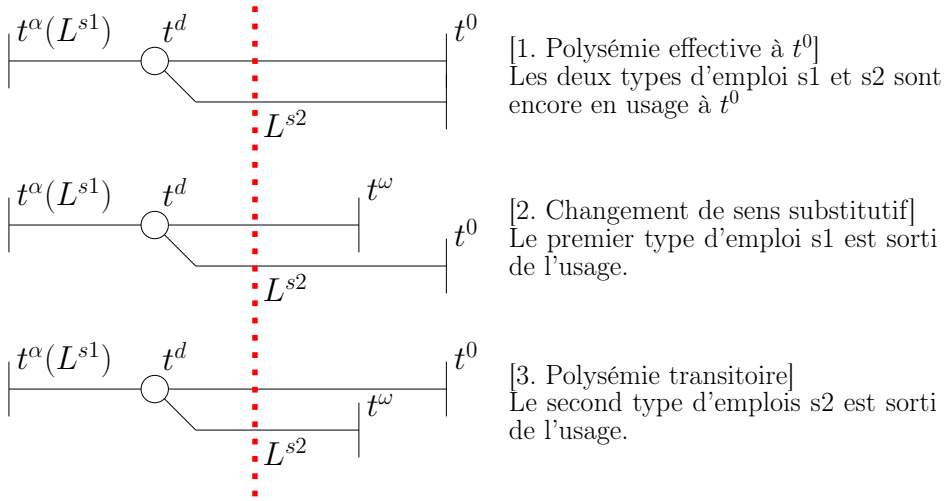


FIGURE 5 – Trois types de polysémie évolutive (cf. François 2008 : 10-12)

Dans la réalité, la question est plus complexe, car il faut tenir compte des technolectes, c.à.d. de vocables dont un ou plusieurs sens sont techniques et n'ont jamais été connus que de spécialistes. Ainsi, dans l'étude par A. Darmesteter (1887 : 81-2), des « changements de sens » du n.m. *timbre*, certains des 12 sens du vocable que l'auteur énumère ne sont connus que des musiciens, ex.

> [5] Caractère d'un son indépendamment de son rang dans l'échelle, caractère tenant à des sons harmoniques qui coexistent avec le son fondamental

à des historiens des techniques, ex.

> [8] Marque particulière que chaque bureau de poste imprime sur les lettres, indiquant le lieu et le jour du départ pour celles qui partent et le lieu et le jour de l'arrivée pour celles qui arrivent.

ou ont complètement disparu de l'usage courant du 21^{e} siècle, ex.

> [6] Premier vers d'un vaudeville connu, qu'on écrit au-dessus d'un vaudeville parodié pour indiquer sur quel air ce dernier doit être chanté.

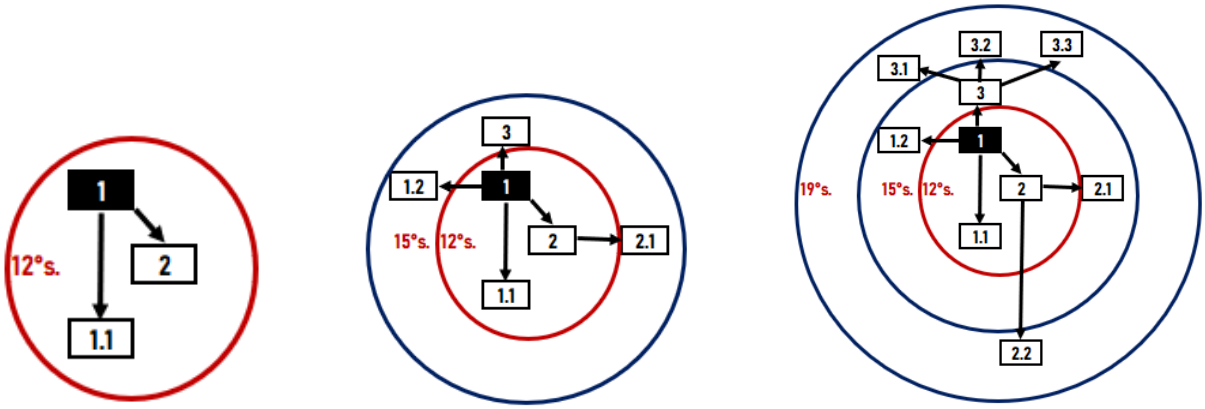
1.2 Deux techniques en concurrence

Dans la phase initiale du projet, Jacques François a testé deux techniques de visualisation[7]. La première (Fig. 6, divisée en 3 'scans') subdivise l'évolution de la polysémie d'un vocable en « instantanés séculaires ». L'évolution de cette polysémie abstraite – c.à.d. sans réalisation attestée – passe par trois périodes[8] :

— le 12e siècle, où le sens premier général [1] apparaît ainsi qu'une dérivation [1.1] et une extension [2]

— le 15e siècle, où une seconde extension [3] émerge, tandis qu'une seconde dérivation du premier sens [1.2] s'ajoute à la première, et que le sens [2] donne lieu à la dérivation [2.1],

— et le 19e siècle, où une seconde dérivation du sens [2], [2.2] et trois dérivations simultanées du sens [3], [3.1/2/3] complètent le processus.



(a) 12e siècle (b) 15e siècle (c) 19e siècle

FIGURE 6 – Visualisation d'une polysémie évolutive abstraite par instantanés séculaires

La seconde technique consiste à construire un graphe historique dont chacun des nœuds symbolise l'un des sens répertoriés dans la fig. 7 dans un cadre orthogonal, les siècles se succédant en abscisses et les sens en ordonnées (orientées de haut en bas, afin de pouvoir être associées à un tableau de commentaires, sens par sens).

L'avantage de la visualisation scannée en instantanés séculaires est de donner une idée précise de l'état d'évolution de la polysémie du vocable à chacun des siècles où une diversification des sens est attestée, c'est le type de visualisation adopté dans François (2020) pour représenter la polysémie évolutive de l'adj. et n.m. *curieux*, cf. Fig. 8

Mais cette technique a un inconvénient : siècle par siècle, les nouveaux sens sont répartis dans l'espace attribué au siècle concerné, mais leur disposition relative n'est pas raisonnée. En outre, si deux diversifications de sens successives ont eu lieu à un ou plusieurs siècles de distance, les siècles d'invariabilité de la polysémie ne figurent pas. De ce fait – et aussi pour économiser de la place – nous n'avons pas encore cherché à automatiser

7. Le mode de visualisation de la polysémie évolutive illustré par la fig. 6 appliqué à l'adj. et n.m. *curieux* (François 2020a : 79-84), au n.f. *terre* (François (2020b : 27, 29), et au n.f. *batterie* (François 2021b : 64) et celui illustré dans la fig. 7 à la polysémie évolutive des v. *atterrer, bouleverser, navrer, meurtrir et prévenir* ainsi que des adj. *formidable* et *frais* et des n.m. *régime* et *rôle* (François 2021 : 65-85). Voir aussi l'évolution comparée de la polysémie du n.f. *terre* et du n.m. *monde* (François 2020b : 36).

8. On ne tient pas compte ici des cas de déclin, voire de disparition de l'un ou l'autre sens, car leur datation est difficile à attester.

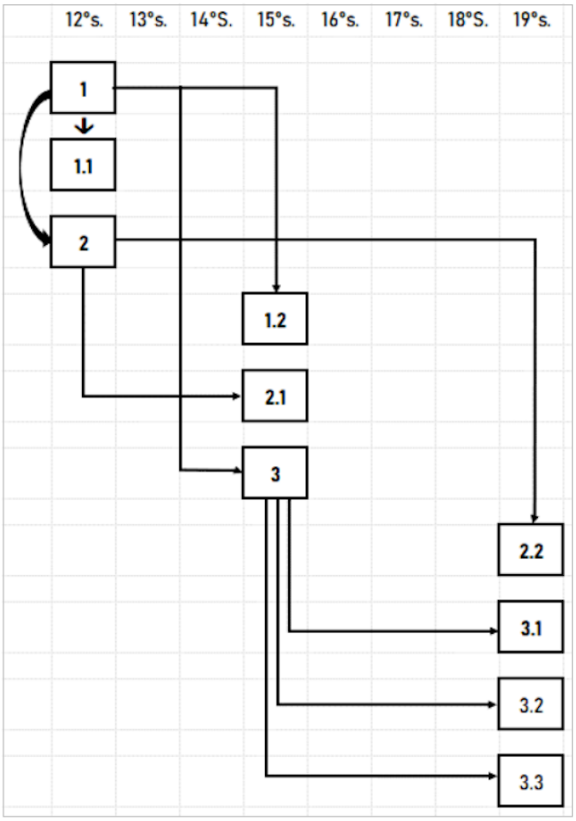
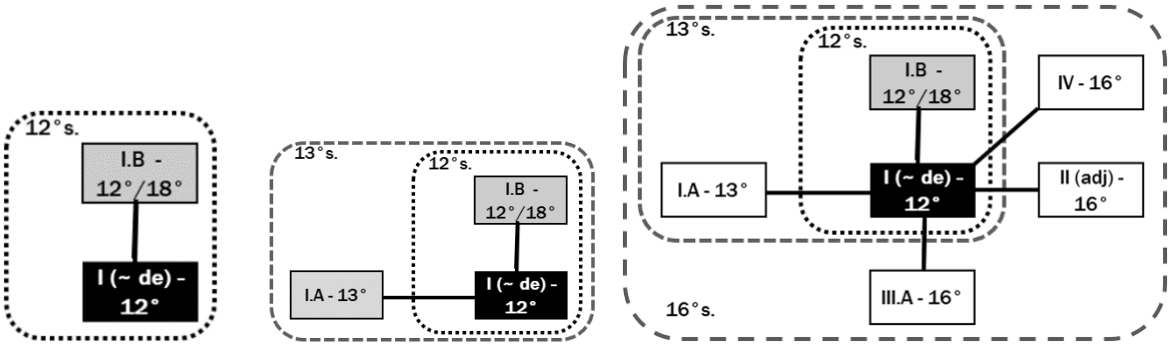


FIGURE 7 – Visualisation de la même polysémie évolutive par un graphe historique situé dans un cadre orthogonal siècles × sens



(a) 12e siècle (b) 15e siècle (c) 19e siècle

Les sens véhiculés par les nœuds des trois diagrammes se laissent paraphraser par :

I.A : qui s'inquiète de (qqc/qqn) ;

I.B : désireux de (qqc) ;

II, adj : désireux de voir, de savoir ;

III.A, subst : indiscret ;

IV, adj : original, extraordinaire, digne d'intérêt.

FIGURE 8 – Fragment de la visualisation de la polysémie évolutive de l'adj. et n.m. curieux par « instantanés séculaires » (Cf. François 2020a : 37-38)

cette technique. En revanche, la visualisation de la polysémie évolutive d'un vocable par un graphe historique du type suggéré dans la figure 7 présente deux avantages :

1. La disposition des sens sur un damier sous-jacent attribue à chaque sens une position associée à son siècle d'apparition (en abscisses) et à son rang (en ordonnées). Comme deux sens ne peuvent pas figurer au même niveau, chacun de ces niveaux peut être associé à une ligne du tableau de commentaires.
2. Et en outre on peut aisément apprécier l'ensemble des sens apparus respectivement au 12e, au 15e et au 19e siècle, puisqu'ils se déclinent sur une même colonne, tandis que les siècles où aucune diversification de sens n'est attestée se distinguent tout aussi clairement, permettant au lecteur de mesurer l'écart chronologique entre les périodes de diversification des sens.

Pour cette double raison, nous privilégions actuellement ce second type de visualisation qui permet entre autres de comparer les profils évolutifs de groupes de vocables présentant des parentés morphologiques et/ou sémantiques. L'une des premières retombées de ce 'profilage' des polysémies évolutives est l'hypothèse (à étayer à partir d'un nombre beaucoup plus conséquent de vocables) que les siècles de diversification majeure des sens lexicaux ont été le 12e, le 16e et le 19e et non par exemple le 17e, en dépit de l'émergence des premiers dictionnaires de langue, notamment celui de Furetière en 1690, phénomène éditorial et culturel majeur qui a nécessairement incité le public lettré (certes insignifiant par rapport à la masse des illettrés, mais influent dans la société des notables) à se confronter à ces références lexicographiques et à prendre conscience de la constitution d'un « bel usage » normatif de la langue française.

La comparaison de ces profils évolutifs – sémantiques ou plutôt sémantaxiques, puisque les nouveaux sens transparaissent contre l'arrière-plan de nouvelles combinaisons syntaxiques privilégiées – laisse supposer que L'ÉVOLUTION DES POLYSÉMIES LEXICALES S'EST EFFECTUÉE PAR VAGUES, entre des périodes d'exploitation exubérante des potentialités d'un état du lexique (on pense inévitablement à la créativité lexicale exceptionnelle de Rabelais) et des périodes de reconfiguration collective du lexique (par Montaigne, Malherbe, les poètes de la Pléiade, etc.).

2 Le projet d'informatisation fonctionnelle et de modélisation graphique des entrées historico-étymologiques [H-É] du TLFi

Les dictionnaires de la langue française – dans son intégralité historique ou limités à des époques délimitées – peuvent entrer dans trois classes (cf. François 2009 pour plus de détails) qui résultent de deux niveaux de distinction, d'abord entre l'approche **synchronique** et l'approche **historique** (ou longitudinale), ensuite parmi les ouvrages décrivant un état de langue, entre les dictionnaires **contemporains** (au sens premier, c'est-à-dire qui décrivent l'état évolutif de la langue que le lexicographe pratique lui-même) et les dictionnaires **philologiques** :

— DICTIONNAIRES SYNCHRONIQUES ET CONTEMPORAINS
Les premiers d'entre eux remontent au 14e siècle avec le *Dictionnaire françois-latin* de Robert Estienne (1539) suivi en 1610 du *Thresor de la Langue Françoyse* de Jean Nicot (dans le même esprit puisque l'auteur se contente d'associer à chaque entrée un équivalent en latin et de traduire également les exemples en latin classique).

Les principaux dictionnaires « de langue » – à distinguer des dictionnaires « de choses » qui prennent leur source dans l'*Encyclopédie* de Diderot et d'Alembert (1751-72) sous-titrée *Dictionnaire raisonné des sciences, des arts et des métiers* – sont mentionnés par siècle dans la colonne de gauche de la fig. 9.

— Dictionnaires synchroniques et philologiques
Dans ces ouvrages, les auteurs sélectionnent un état de langue passé. Le dictionnaire de l'ancien français de Frédéric Godefroy (1881-1902) est le plus ancien et couvre également le moyen français, tout comme le *Altfranzösisches Wörterbuch* de Tobler et Lommatzsch, dont la publication en fascicules par Adolf Tobler, puis Erhard Lommatzsch et enfin Hans Helmut Christmann, a pris près d'un siècle avant que le dictionnaire bénéficie en 2002 d'une édition électronique sur CD-ROM par Peter Blumenthal et Achim Stein (université de Stuttgart). Quant au *Dictionnaire du Moyen Français* conçu par Robert Martin et développé à l'ATILF sous la direction de Sylvie Bazin, il a un format électronique et est accessible par le site du CNRTL (ou par LEXILOGOS [9]).

— Dictionnaires étymologiques et dictionnaires historiques
Enfin, bien qu'ils aient deux objets d'étude apparentés, les dictionnaires historiques (expression d'une lexicographie longitudinale) se distinguent des dictionnaires étymologiques, car leur propos est de décrire l'évolution des usages des vocables, leur variation phonétique et orthographique, syntaxique (notamment l'apparition et le déclin de collocations et de locutions figées) et sémantique (leur place dans des champs sémantiques appelés à accueillir, déplacer et exclure au fil des siècles les vocables en concurrence permanente). Les articles du dictionnaire d'Émile Littré qui concernent des vocables dont l'origine est antérieure au 17e siècle ont une section historique. À titre d'exemple, celle – particulièrement copieuse – du n.m. ***coeur*** énumère 137 citations qui se répartissent ainsi :

11e	12e	13e	14e	15e	16e
5	30	23	6	28	45

Mais ces citations classées par siècle et ne faisant l'objet d'aucun commentaire sémantique, ne constituent qu'un aide-mémoire peu instructif des usages du vocable jusqu'au 16e siècle. De leur côté, les entrées historiques du TLFi – au nombre de **49 876** selon notre dernière évaluation, dont **20 843** ont une microstructure révélant la polysémie évolutive du vocable examiné – classent leurs rubriques selon les types d'usage successifs des vocables du 9e au 20e siècle en mettant en valeur la forme des mots et leurs collocations. Le seul ouvrage entièrement consacré à l'histoire du lexique français est le *Dictionnaire historique de la langue française* [10], dont les entrées bénéficient des recherches des éditions Le Robert complétées par de nombreux emprunts au TLFi.

Le diagramme de synthèse historique en fig. 9 réunit deux stades historiques :

— d'abord les principaux dictionnaires de la langue française depuis leur apparition au début du 17e siècle et les dictionnaire spécialisés, étymologiques ou historiques, jusqu'à la finition du TLF,

— ensuite le TLF avec ses deux éditions successives, en 16 volumes et sur support électronique, avec ses deux principaux types d'entrée [11]

Et le diagramme en fig. 10 indique le positionnement de notre double projet dans le sillage du TLF dans son édition originale et du TLFi.

9. cf. https://www.lexilogos.com/francais_dictionnaire.htm
10. Le DHLF est paru en 1992, sous la direction d'Alain Rey, et vient d'être réédité en 2022 pour son

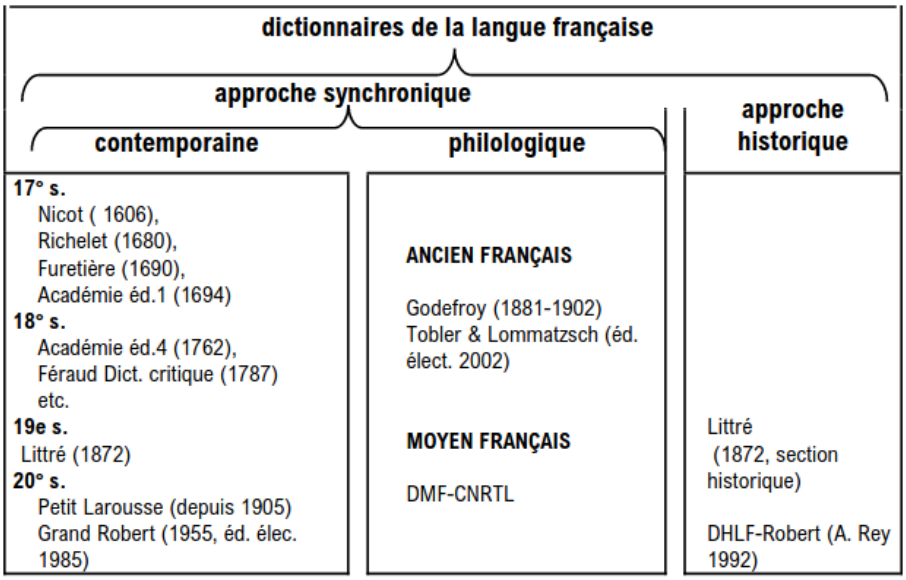


FIGURE 9 – Classement des dictionnaires antérieurs au TLF et dont il a tiré parti

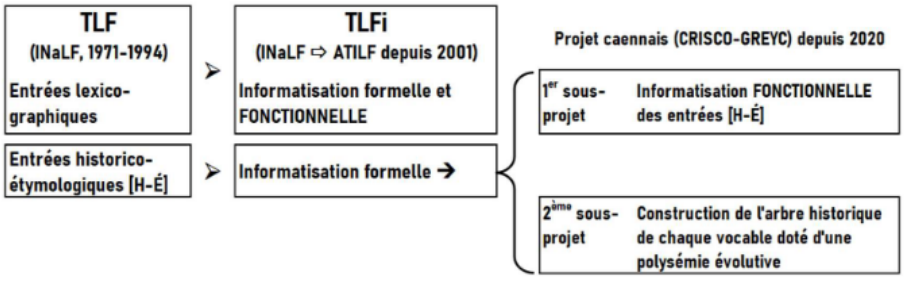


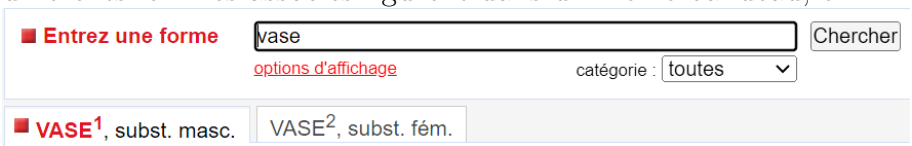
FIGURE 10 – Positionnement de notre projet dans le prolongement de l'informatisation du TLF

3 La microstructure des articles du TLFi et de leur entrée H-É

3.1 Articles, entrées, dégroupements et regroupements

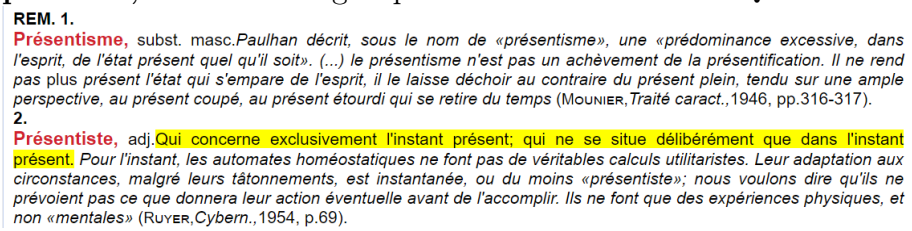
Dans une présentation des caractéristiques techniques du TLFi et de leur exploitation parue en 2010, Pascale Bernard, de l'ATILF, assimile [12] *mots traités en vedette d'articles* et *entrées principales*. Comme dans le *Dictionnaire du Français Contemporain* (J. Dubois, dir. Larousse 1966) et le *Dictionnaire LEXIS de la Langue Française* (J. Dubois, dir., Larousse 1975, rééd. 2002), la notion d'article dictionnairique est liée à une double pratique lexicographique :

1. D'un côté, l'inventaire des entrées lexicographiques est **supérieur** à la liste des lemmes figurant dans le dictionnaire en raison de **dégroupements** (dans ce cas, les lemmes dégroupés ont une forme identique accompagnée d'une numérotation en exposants). On peut considérer que la partie commune définit un article puisque les différents lemmes associés figurent dans un même bandeau, ex.



   Le critère du dégroupement peut être étymologique, ex. **dé[1]** (à jouer) *vs* **dé[2]** (à coudre), morphologique et/ou sémantique, ex. **présent[1]**, -ente (adj.) *vs* **présent[2]** (n.m. temps présent) *vs* **présent[3]** (n.m. cadeau).

2. D'un autre côté, l'inventaire des entrées lexicographiques est inférieur à la liste des lemmes en raison de **regroupements** morpho-sémantiques. Ainsi dans l'entrée **présent[1], -ente** sont regroupés sous l'intitulé REMARQUES :


REM. 1.
Présentisme, subst. masc. *Paulhan décrit, sous le nom de «présentisme», une «prédominance excessive, dans l'esprit, de l'état présent quel qu'il soit». (...) le présentisme n'est pas un achèvement de la présentification. Il ne rend pas* plus *présent l'état qui s'empare de l'esprit, il le laisse déchoir au contraire du présent plein, tendu sur une ample perspective, au présent coupé, au présent étourdi qui se retire du temps* (MOUNIER, *Traité caract.*, 1946, pp.316-317).
2.
Présentiste, adj. *Qui concerne exclusivement l'instant présent; qui ne se situe délibérément que dans l'instant présent. Pour l'instant, les automates homéostatiques ne font pas de véritables calculs utilitaristes. Leur adaptation aux circonstances, malgré leurs tâtonnements, est instantanée, ou du moins «présentiste»; nous voulons dire qu'ils ne prévoient pas ce que donnera leur action éventuelle avant de l'accomplir. Ils ne font que des expériences physiques, et non «mentales»* (RUYER, *Cybern.*, 1954, p.69).

L'article se décompose entre une entrée principale, dite LEXICOGRAPHIQUE, suivie d'une rubrique PRONONCIATION ET ORTHOGRAPHE, d'une entrée historico-étymologique (désormais abrégées « H-É ») et généralement de deux rubriques, STATISTIQUES et BIBLIOGRAPHIE. Dans le cas d'un article dégroupé pour une raison autre qu'étymologique (ex. ÉCHEC 1/2), une seule entrée historico-étymologique est attachée à l'une des entrées lexicographiques, dans le cas contraire chaque entrée lexicographique donne lieu à une entrée historico-étymologique distincte. Les entrées H-É (encore intitulées ÉTYMOLOGIE ET HISTOIRE malgré l'ordre inversé au-delà du premier volume imprimé du TLF) [13] sont implicitement subdivi-

30ème anniversaire.

11. Les articles du TLFi comportent également une entrée phonétique, des références bibliographiques et généralement une entrée statistique (sur la fréquence du vocable examiné aux 19e et 20e siècle sur la base des textes enregistrés dans FRANTEXT et une entrée dérivationnelle pour des vocables dénués d'article propre.

12. Cf. P. Bernard (2010, § 13 de la version HTML) : « Les listes défilantes permettent de parcourir le TLF, comme on pourrait le faire dans la version papier, mais elles affichent uniquement les mots traités en vedettes d'articles, c'est-à-dire en entrées principales ».

13. Cette disposition régulière ne vaut pas encore au tout début du dictionnaire. Dans un premier temps l'entrée est extrêmement développée, les deux sous-entrées sont clairement distinctes et la place attribuée à la Révolution française dans l'histoire du français est soulignée avec deux intertitres la prenant

sées en une sous-entrée historique immédiatement suivie d'une sous-entrée étymologique. Ces entrées H-É se laissent classer en deux types :

1. celles qui sont subdivisées en rubriques introduites par un identifiant hiérarchisé commençant par un caractère du type I, A, 1 ou a (tout identifiant a un format simple ou hiérarchisé et dans ce dernier cas, un 5ème rang α peut intervenir) et qui sont donc dotées d'une POLYSÉMIE ÉVOLUTIVE,
2. et celles qui ne sont pas divisées en rubriques et qui sont donc dénuées de polysémie évolutive selon le TLFi. À titre d'illustration, nous allons explorer les deux entrées, lexicographique et historico-étymologique, de l'article du n.m. CŒUR, l'un des plus développés.

3.2 Comparaison entre les deux entrées lexicographiques du H-É du n. m. CŒUR

L'entrée lexicographique de CŒUR a un volume de 13 466 mots graphiques, elle contient 69 citations littéraires réparties entre 2 classes superordonnées I et II, déclinées respectivement en deux sous-classes A-B et quatre sous-classes A-B-C-D (cf. Tab. 1 qui ne mentionne que les deux premiers rangs hiérarchiques du classement des rubriques). Au total, le classement hiérarchique met en œuvre les quatre premiers niveaux : I, A, 1 et a.

I	[Le cœur dans sa réalité physique]
	A [Le cœur comme organe interne]
	B P. *méton.* Poitrine, qui abrite le cœur (et, secondairement, les autres organes internes primordiaux) ; en *partic.* l'endroit de la poitrine où les battements du cœur sont perceptibles
II	[Le cœur comme foyer ou réceptacle de la vie intérieure] Qui ne sait qu'une physiologie peu exercée a donné au cœur un rôle, peu défini, mais excessif, comme organe de toute notre vie intime ? (*Théol. cath.* t. 3, 11911).
	A [P. réf. à l'automatisme cardiaque ; le cœur comme organe ou lieu d'une saisie plus ou moins automatique]
	B [P. réf. à l'intériorité et à l'activité de l'organe en tant que facteur central de la vie hum. individuelle] *Le cœur comme foyer ou réceptacle de la vie intérieure profonde, de la personnalité morale d'un individu*.
	C [Le cœur comme foyer ou réceptacle du dynamisme moral, de certaines tendances volitives]
	D [Le cœur comme foyer ou réceptacle de la vie affective]

TABLEAU 1 – Les deux premiers niveaux hiérarchiques de la microstructure de l'entrée lexicographique du n.m. CŒUR. Le surlignage est effectué par l'ATILF à partir de l'information fonctionnelle de l'entrée lexicographique.

.

L'entrée H-É a un volume de 817 mots graphiques. Elle se subdivise en deux classes superordonnées I et II, lesquelles se déclinent respectivement en huit rubriques pour I et

en compte, par ex. dans l'article ABSTENIR (S'), partie Histoire, « I.– Disparitions av. 1789 » vs « II.– Hist. des sens et emplois attestés apr. 1789 ». La présence d'entrées historiques et étymologiques « hors normes » au tout début du dictionnaire et l'absence de démarcation régulière entre les deux sous-entrées historique et étymologique dans la suite des entrées H-É, constituent deux obstacles difficiles à contourner pour l'informatisation fonctionnelle de ces entrées.

quatre pour II. Ces douze rubriques se décomposent à leur tour en 35 sections, soit une moyenne de trois sections par rubriques. La hiérarchie des rubriques est modérée, elle ne met en œuvre que les deux premiers rangs d'identifiants : I à II et A à H. Dans le tableau 3.2 ci-dessous, nous avons surligné les données qui se laissent catégoriser fonctionnellement, selon le modèle de l'ATILF :

— vert : prédéfinition (sans guillemets)

— gris : forme de mot / collocation

— jaune : définition (avec guillemets) / quasi-définition (sans guillemets)

— bleu : citation dans une référence (entre parenthèses).

La sous-entrée étymologique figure encadrée sur fond gris (cf tableau 3.2 page 18).

Dans l'entrée lexicographique, le surlignage de l'ATILF a été rendu possible par l'informatisation fonctionnelle du texte dans la version papier originelle. Quant à notre surlignage analogue de l'entrée H-É, il résulte de quatre constats typographiques :

— les chaines de caractères (avec ou sans espace blanc interne) en caractères italiques hors parenthèses véhiculent généralement la **forme** du mot à l'époque concernée ou une **collocation** ;

— celles entre guillemets véhiculent une **définition** ;

— celles introduites par < : > dans un espace entre parenthèses et comportant la vedette en caractères gras véhiculent des **citations** ;

— celles en **caractères romains** figurant immédiatement après un identifiant de classe superordonnée, véhiculent une « pré-définition ».

En même temps, cet exercice visant à préfigurer l'informatisation fonctionnelle des entrées H-É, fait ressortir des obstacles qu'il importe de contourner dans cette entreprise, notamment :

— la disparité des modes de **datation** de la première attestation d'un vocable dans l'emploi spécifié par l'identifiant de la rubrique, qui incite à remplacer tous ces modes de datation épars par la mention du siècle, ce qui débouche sur une échelle chronologique de onze siècles entre le 9^{e} (celui des premiers manuscrits enregistrés dans FRANTEXT) et le 20^{e} (le dernier siècle entier) ;

— l'absence de séparation claire entre la sous-entrée **historique** et la sous-entrée **étymologique**, qui oblige à explorer des centaines de débuts d'étymologie, afin d'exploiter les plus fréquents pour un traitement automatisé, inévitablement suivi d'un post-traitement manuel ;

— et la difficile attribution d'une fonction lexicographique précise à de nombreux segments figurant en caractères romains (par défaut) et dont seuls la disposition dans la **série linéaire** des données et l'éventuelle signalisation comme une **abréviation** peuvent entrer en ligne de compte, s'il s'avère souhaitable de les enregistrer distinctivement dans l'informatisation fonctionnelle. Cela concerne entre autres les 'quasi-définitions' comme dans le Tab. 2 (rubrique II.D) avec le renvoi à une définition figurant dans le *Dictionnaire universel* de Furetière en 1690 :

> « 1690 (Fur. : On dit aussi, qu'on fait dîner quelqu'un **par cœur** quand on ne luy a point donné à dîner) ». Repérer automatiquement de telles quasi-définitions implique le recours à une « expression régulière » recherchant à gauche un segment en petites capitales suivi d'un point [abréviation du nom d'un auteur de dictionnaire] et du signe typographique < :>, et à droite une parenthèse fermante.

I. Organe central de la circulation.

[I.]A. *Ca* 1050 « siège de la vie » (*Alexis*, éd. Ch. Storey, 445 :Ço'st granz merveile que li mens quors tant duret) ;
1130-40 (Wace, *Ste Marguerite*, éd. E. A. Francis, 62 : Ele ama Deu et Deu l'ama, Trestot son cuer li adona).

[I.]B. *Ca* 1100 au propre (*Roland*, éd. J. Bédier, 2965 : [Li emperere ad fait] tuz les quers en paile recuillir).

[I.]C. *Ca* 1100 p. ext. « la poitrine » (ibid., 3448 : L'escut li freint, cuntre le coer li quasset).

[I.]D. 1195-1200 « siège des sensations physiques » (*Renart*, éd. Martin, branche 11, 565 : il avoit a son **cuer** grant fein) ;
ca 1200 « estomac » (*ibid.*, branche 9, 1724 : A pou que li **cuers** ne me faut) ;
13^{e}s. « région épigastrique » (J. Le Marchand, *Mir. N.-D. de Chartres*, 5 ds T.-L. : a vomir les convenoit Du mal qui au **cuer** leur venoit) ;
1508 *dire tout ce qu'on a sur le cœur* (Eloy d'Amerval, *Livre de la Deablerie*, 147b ds *IGLF*) ;
1633 *coucher du cœur sur le carreau* « vomir [jeu de mots tiré des cartes] » (*Comédie des Proverbes*, acte II, scène 2, Anc. Théâtre fr., t. 9, p. 42).

[I.]E. Fin 12^{e}s. « partie centrale » (*Mort Aymeri de Narbonne*, 607 ds T.-L. : El **cuer** de France).

[I.]F. 1340 « objet en forme de cœur » (v. Gay).

[I.]G. 1585 « as de cœur » (N. Du Fail, *Contes d'Eutrapel*, t. 2, p. 202 ds *IGLF*).

[I.]H. 1600 « sorte de cerise » (Ol. de Serres, *Théâtre d'Agric.*, VI, 26 ds Hug.).

II. Centre de la vie intérieure.

[II.]A. Siège des émotions, de l'affectivité. *Ca* 1050 (*Alexis*, 464 : Ne puis tant faire que mes **quors** s'en sazit) ;
ca 1100 (*Roland*, 317 : Tro avez tendre **coer**) ;
1ertiers 13^{e}s. (*Lancelot du Lac*, éd. O. Sommer, t. 5, partie 3, p. 353 : il navoit oi noveles ... qui tant li feissent mal au **cuer**) ;
1167-70 p. méton. *cœur* désigne la personne chérie (G. d'Arras, *Ille et Galeron*, 4160 ds T.-L.).

[II.]B. Siège du désir, de la volonté.
Ca 1050 (*Alexis*, 166 : Quant tut sun **quor** en ad si afermét) ;
ca 1162 *de son cuer* « de toute son ardeur, très sincèrement » (*Flore et Blancheflor*, 1925 ds T.-L.) ;
début 14^{e}s. *avoir au cuer de* (faire qqc.) (*Ovide moralisé*, éd. C. de Boer, livre V, 460) ;
1579 *de gayeté de cœur* (H. Estienne, *Precellence du lang. fr.*, 359 ds *IGLF*) ; 1585 *du meilleur de mon cœur* (N. Du Fail, *Contes d'Eutrapel*, t. 2, p. 275). 1586 *à cœur vaillant, rien impossible* (E. D'Amerval, *loc. cit.*, 138b).

[II.]C. Siège de l'intelligence.
1130-40 « discernement » (Wace, *Ste Marguerite*, 431 : Lors **cuers**, lor sens, fais oscurer) ;
ca 1190 « *savoir intuitif* » (*M. de France, Lais, Guigemar*, 547, éd. J. Rychner : Mis **quors** me dit que jeo vus pert) ;
ca 1220 les ielz dou cuer (G. de Coincy, *Mir.*, éd. Koenig, II Ch 9,3792) ; *cf.* au 17^{e}s. le cœur en tant que siège de la grâce, permettant la communication avec Dieu (Pascal, *Pensées*, section IV, 278 et 277, éd. Brunschvicg, t. 13, p. 201 : C'est le **cœur** qui sent Dieu, et non la raison ; le **cœur** a ses raisons, que la raison ne connaît point ; section XII, 793, t. 14, p. 232 : aux yeux du **cœur** et qui voient la sagesse).

[II.]D. Siège du souvenir, de la mémoire.
Ca 1190 (M. de France, *Fables*, 70, 61 ds T.-L. : Senz quer fu e senz remembrance) ;
ca 1200 retenir par cuer (*Poème moral*, éd. Bayot, 1036) ;
ca 1220 savoir par cuer (G. de Coincy, *Mir.*, éd. Koenig, I Mir 11, 757) ;
1690 (Fur. : On dit aussi, qu'on fait dîner quelqu'un par cœur quand on ne luy a point donné à dîner) ;
1694 p. ext. *de savoir par cœur* : *apprendre une chose par cœur* (Ac.), v. aussi Tobler, *Sitzung der philosophisch-historischen Classe vom 27. October 1904*, Berlin, p. 1274, 1275.

Du lat. class. *còr* (peut-être par l'intermédiaire d'une forme **còre*, Fouché, p. 656, Bl.-W.$^{1\text{-}5}$) qui, dans la conception antique, est à la fois le siège de la vie et des fonctions vitales, et celui des passions et des émotions, des pensées et de l'intelligence, de la mémoire et de la volonté (cf. gr. $\kappa\alpha\rho\delta\iota\alpha$ « cœur » et aussi « entrée de l'estomac », « siège des passions et des facultés de l'âme » ; v. aussi K. Weinberg ds *Arch. St. n. Spr.*, t. 203, 1966-67, pp. 1-31) ; pour par *cœur*, v. Bambeck, *Lat. rom. Wortstudien*, n°126.

Tableau 2 – L'entrée H-É du n.m. CŒUR (déployée visuellement en attribuant une ligne à chaque section et à chaque rubrique dénuée de section)

4 L'éventail typographique des entrées H-É

L'informatisation fonctionnelle des entrées H-É peut s'effectuer pas à pas à l'aide de différents outils dont seule une combinaison ingénieuse (et progressivement experte) permet de se rapprocher du but. La première méthode consiste à tirer parti de certains des signes typographiques figurant dans les entrées. Ces signes se laissent subdiviser en deux catégories, les simples et les doubles. Leur fonction respective est inverse :

— les signes simples, notamment le point-virgule et le point, ont une fonction de **séparation** entre deux espaces à gauche et à droite,

— tandis que les signes doubles "ou discontinus" (guillemets-chevrons « ... », parenthèses courbes et parenthèses droites) ont une fonction inverse d'**encadrement** d'une suite de chaînes de caractères.

En outre, l'interprétation fonctionnelle des signes typographiques est souvent dépendante de la police dans laquelle se présentent les chaînes de caractères figurant immédiatement avant ou après un signe simple, ou entre les deux éléments d'un signe double.

Dans les entrées H-É, la fonction du point-virgule (suivi d'un espace vide) est de séparer :

— soit la dernière section d'une rubrique, de l'identifiant de la rubrique suivante qui figure en caractères gras,

— soit deux sections successives d'une même rubrique, si la chaîne de caractères immédiatement à droite figure dans un autre type de caractères.

De son côté, la fonction du point varie essentiellement selon quatre types de configurations :

1. La seule chaîne de caractères immédiatement à gauche figure en gras. Dans ce cas le point clôt un identifiant, il est suivi d'un espace vide et d'un caractère alphabétique en majuscules ou d'une date [14], ex. (entrée PARTICULARISME, aspect de l'édition publique et notation XML [15]) :

Étymol. et Hist.1. 1689 théol.	`<etymology><hi rend="bold">Étymol. et Hist.1.</hi> 1689 théol. (...)`

2. La seule chaîne de caractères immédiatement à droite figure en gras. Dans ce cas, le point sépare la fin d'une rubrique de 1er rang (dont le premier composant de l'identifiant est un chiffre romain, ex. `I.`) de l'identifiant (en gras) de la rubrique suivante, toujours de premier rang, ex. `II.` (cf. article PARTICULIER, -IÈRE), ex.

II. Subst. **A.** masc	`(...) <hi rend="bold">II.</hi> (...)`

3. Ou les deux chaînes de caractères situées immédiatement à gauche et à droite figurent en gras. Dans ce cas, le point sépare deux composants d'un identifiant avec en général un espace vide entre ceux-ci, ex. (article PARTITION)

PARTITION, subst. fém. **Étymol. et Hist.I. 1.** *Ca* 1175	`<hi rend="bold">Étymol. et Hist.I. 1.</hi>`

14. Si le dernier composant de l'identifiant est un caractère alphabétique en minuscules (a, b, etc.), ce caractère est précédé d'un espace vide et le point est remplacé par une parenthèse courbe fermante

15. Du point de vue fonctionnel, une balise </hi> devrait annuler le type de caractère gras immédiatement après « Étymol. et Hist. » pour clore le segment de titre de l'entrée historico-étymologique et une nouvelle balise <hi rend="bold"> devrait ouvrir le segment de l'identifiant de la première rubrique. Mais du point de vue formel, cette opération est inutile, puisque la différence de fonction attribuée au type de caractère gras dans les deux segments successifs n'est pas prise en compte à ce niveau.

4. Ou encore aucune des deux chaînes de caractères situées immédiatement à gauche et à droite ne figure en gras. Dans ce dernier cas, le point, suivi d'un espace vide, clôt une abréviation (ex. « Part. » pour « Participe », article PARVENU, -UE, adj. et subst.), ex.

Part. passé de *parvenir**.	`Part. passé de <hi rend="italic">parvenir*</hi>`

La fonction des guillemets est de distinguer une définition, mais il est à noter qu'on rencontre occasionnellement des 'quasi-définitions', introduites par un verbe de désignation mais dénuées de guillemets, ex. (rubrique A.2 de l'entrée PAVOIS) :

> **2.** (...) ; 1874 désigne les pavillons dont on orne le gréement d'un navire en signe de réjouissance

La fonction des parenthèses courbes est d'enregistrer des références, y compris éventuellement une citation (introduite par " :" et en caractères romains), ex. (entrée PAYE, PAIE, rubrique 2) :

2. 1176-81 «action de payer» ici au fig. (CHRÉTIEN DE TROYES, *Chevalier Charrette*, éd. M. Roques, 6893) ;	`2.</hi>1176-81 «action de payer» ici au fig. (<hi rend="smallcaps">Chrétien de Troyes, </hi><hi rend="italic">Chevalier Charrette</hi>, éd. M. Roques, 6893)`

sauf si elles se présentent dans un espace délimité par des guillemets, auquel cas elles introduisent une information secondaire dans une définition, ex.

g) 1640 absol. «acquitter un droit (en parlant de marchandises)»	`<hi rend="bold">g) </hi>1640 absol. «acquitter un droit (en parlant de marchandises)»`

Quant à la fonction des parenthèses droites [...], elle est de fournir des informations présentées sous réserves, notamment en matière de manuscrits, ex. entrée PASSER, rubrique A.7 :

7. [fin XIII^e^s. soi passer pour «être considéré comme» (Sone de Nansai, 20478 ds T.-L.)]	`<hi rend="bold">7.</hi> [fin <hi rend="smallcaps">xiii</hi><R/>es. <hi rend="italic">soi passer pour</hi> « être considéré comme» (<hi rend="italic">Sone de Nansai</hi>, 20478 ds T.-L.])`

4.1 La fonction variable des types de caractères selon qu'ils figurent entre ou en dehors de parenthèses courbes

Les rédacteurs des entrées H-É du TLF ont attribué aux caractères gras, italiques et petites capitales des traits fonctionnels différents selon leur encadrement typographique (table 3).

Ainsi par exemple la mise en forme de la rubrique ci-dessous :

> **2.** indique un changement d'état, d'aspect **a)** *ca* 1165 *torner a porreture* (en parlant d'un cadavre) (BENOÎT DE STE-MAURE, *op. cit.*, 22397)

met en œuvre dix segments en caractères romains, gras et italiques avec deux espaces entre parenthèses courbes. Le tableau 4 explicite la fonction de chacun de ces segments.

Les constats effectués dans cette section sont convertibles en « expressions régulières » opérant à partir [1] du repérage des signes typographiques simples et doubles et [2] des

	En dehors de parenthèses courbes	**Entre parenthèses courbes**
Caractères gras	Identifiants	Vedette dans une citation
Caractères italiques	Forme de mot ou collocation	Titre d'une référence
Petites capitales	Numérotation des siècles en chiffres romains	Auteur d'une référence

Tableau 3 – Fonction contextuelle des types de caractère entre ou en dehors de parenthèses courbes

Format public	**Type de caractère et parenthèse**	**Format XML**	**Fonction lexicographique**
2.	gras	`<hi rend="bold"> 2. </hi>`	identifiant
indique un changement d'état, d'aspect	romain	`indique un changement d'état, d'aspect`	quasi-définition
a)	gras	`<hi rend="bold"> a) </hi>`	Identifiant (sans rappel du 2.)
ca	italique	`<hi rend="italic"> Ca </hi>`	circa
1165	romain	`1165`	année
torner a porreture	italique	`<hi rend="italic"> torner a porreture </hi>`	collocation
(en parlant d'un cadavre)	(…) & romain	`(en parlant d'un cadavre)`	commentaire
(Benoît de Ste-Maure,	(…) & petites capitales	`(<hi rend="smallcaps">Benoît de</hi> <hi rend="smallcaps">Ste</hi>-<hi rend="smallcaps">Maure</hi>,`	référence : auteur
op. cit.,	italique	`<hi rend="italic"> op. cit. </hi>`	référence : ouvr.
22397)	romain	`22397)`	référence : folio

Tableau 4 – Fonction des types de caractères entre ou en dehors de parenthèses

balises ouvrant et fermant les trois types de caractères employés, en dehors des caractères romains par défaut. Mais attribuer une fonction, même contextuelle, aux segments figurant en caractères romains, est une tâche beaucoup plus délicate. Au stade actuel du projet, il faut distinguer entre les données susceptibles d'être extraites de segments de ce dernier type qui présentent un intérêt majeur pour le tableau de commentaires des nœuds du graphe historique de chaque vocable (section 5.3), et qui méritent de figurer dans une colonne particulière de ce tableau, et celles qu'il suffit – au moins provisoirement – de faire figurer dans la colonne « histoire » sans identification particulière. Les premières se limitent essentiellement à cinq types de métadonnées qui figurent en général en début de rubrique, immédiatement après l'identifiant (tableau 5).

	Classe	**Exemples**
Métadonnées	thématiques	mar. ; tec. ; astr.
	grammaticales	adj. ; subs.
	sémantiques	fig. ; p. ext.
	dialectales	norm. ; pic. ; poit.
	stylistiques (registre)	pop. ; arg.

Tableau 5 – Métadonnées

4.2 La dissociation des deux sous-entrées historique et étymologique

Dans la version publique des entrées H-É (accessibles séparément dans la présentation du CNRTL), il est souvent difficile de repérer le début de la sous-entrée étymologique. Elle est introduite par un point, mais il en est de même pour les identifiants dont le premier composant est de rang I. Au cours de l'élaboration de ces entrées, une configuration particulière est apparue, avec comme séparateur un point suivi de deux espaces vides, voire occasionnellement jusqu'à quatre espaces vides. En tout état de cause, l'écart entre le point et la majuscule qui débute la sous-entrée étymologique est variable, et il faut recourir à une seconde procédure, à savoir le repérage des chaînes de caractères par lesquelles débutent la majorité des sous-entrées étymologiques que la configuration point + 2 espaces vides + majuscule permet déjà d'enregistrer.

Comme précédemment pour les métadonnées des cinq types distingués dans le tableau 5, l'exploitation statistique de la version XML des entrées H-É permet d'extraire et de classer par fréquence décroissante les débuts de sous-entrées étymologiques déjà enregistrées et de les rechercher dans les autres entrées en attente de division. Progressivement, cette opération actuellement en chantier permettra notamment de limiter le contenu de la colonne « forme de mot / collocation » du tableau de commentaires du graphe historique aux segments en italiques (hors parenthèses) de la sous-entrée historique et d'écarter ceux de la sous-entrée étymologique.

5 Le traitement informatique des entrées H-É au format XML

5.1 Structure générale des fichiers

Nous avons vu dans les deux sections précédentes quelles informations nous souhaitons extraire des fichiers XML en précisant quelques critères d'extraction. Dans cette partie, nous allons présenter la structure générale des fichiers XML, développer l'algorithme général et entrer dans le détail des traitements d'extraction implémentés pour obtenir les informations souhaitées.

5.1.1 Les principales balises XML

La version XML des entrées H-É sur lesquelles nous avons travaillé, composée de 81 fichiers, correspond à l'onglet Étymologie de la page web du TLFi [16] comme illustré dans la figure 11 avec l'entrée *lentille* (plante légumineuse, tache de rousseur). Cette présentation met en avant la structure du contenu grâce à la mise en forme, notamment avec les caractères en gras : `Étymol. et Hist. - 1. a) - b) - 2 - 3.`

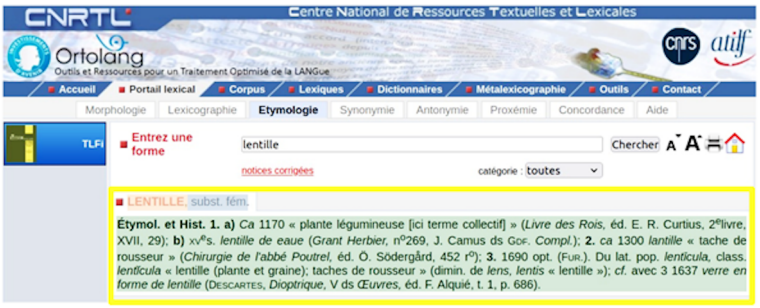


FIGURE 11 – L'affichage public de l'onglet Étymologie de l'entrée LENTILLE

Cette présentation correspond à une *vue*, mise en forme automatiquement, de données XML. La figure 12 montre le code correspondant à l'entrée *lentille*. Comme pour tout document XML, on distingue des balises (éléments entre chevrons), qui ont parfois des attributs (par exemple `rend="bold"`). Les balises délimitent des zones dans le contenu, et peuvent être imbriquées les unes dans les autres, formant une structure arborescente, comme on peut le voir dans la figure 12. Nous pouvons remarquer que les informations sur l'entrée `lentille` sont contenues dans une balise `<entry>` qui contient elle-même trois balises :

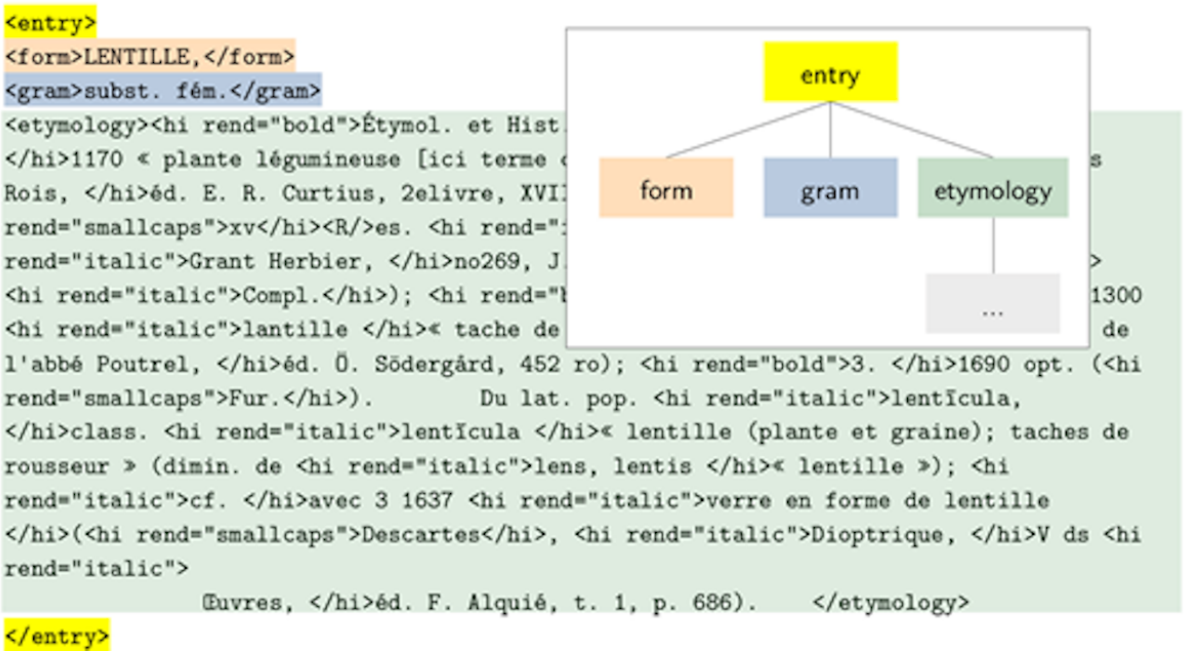
`<form> ... </form>` contient le nom de l'entrée en majuscules suivi d'une virgule ;

`<gram> ... </gram>` contient la catégorie grammaticale de l'entrée ;

`<etymology> ... </etymology>` contient le reste du texte.

Cette structuration est la même pour toutes les entrées du TLFi. Nous voyons que les trois balises `<entry>` puis `<form>` et `<gram>` nous donnent des informations fonctionnelles. La première balise annonce une nouvelle entrée, la seconde l'intitulé de cette entrée et la troisième sa catégorie grammaticale. Enfin, la balise `<etymology>` contient des informations sur l'étymologie du mot. Elle nous fournit la quasi intégralité des informations qui nous intéressent. Cependant, son traitement est plus compliqué comme nous allons le voir ci-dessous.

16. https://www.cnrtl.fr/etymologie/




```
<entry>
<form>LENTILLE,</form>
<gram>subst. fém.</gram>
<etymology><hi rend="bold">Étymol. et Hist
</hi>1170 « plante légumineuse [ici terme
Rois, </hi>éd. E. R. Curtius, 2elivre, XVI
rend="smallcaps">xv</hi><R/>es. <hi rend="
rend="italic">Grant Herbier, </hi>no269, J
<hi rend="italic">Compl.</hi>); <hi rend="
<hi rend="italic">lantille </hi>« tache de
l'abbé Poutrel, </hi>éd. Ö. Södergård, 452 ro); <hi rend="bold">3. </hi>1690 opt. (<hi
rend="smallcaps">Fur.</hi>).        Du lat. pop. <hi rend="italic">lentĭcula,
</hi>class. <hi rend="italic">lentĭcula </hi>« lentille (plante et graine); taches de
rousseur » (dimin. de <hi rend="italic">lens, lentis </hi>« lentille »); <hi
rend="italic">cf. </hi>avec 3 1637 <hi rend="italic">verre en forme de lentille
</hi>(<hi rend="smallcaps">Descartes</hi>, <hi rend="italic">Dioptrique, </hi>V ds <hi
rend="italic">
                Œuvres, </hi>éd. F. Alquié, t. 1, p. 686).    </etymology>
</entry>
```

FIGURE 12 – Code XML – les trois balises principales pour chaque entrée et leur surlignage distinctif.

```
<hi rend="bold">Étymol. et Hist.  1.  a)  </hi>
<hi rend="italic">Ca </hi>
1170 « plante légumineuse [ici terme collectif] » (
<hi rend="italic">Livre des Rois, </hi>
éd. E. R. Curtius, 2elivre, XVII, 29);
<hi rend="bold">b) </hi>
<R/>
<hi rend="smallcaps">xv</hi>
<R/>es.
<hi rend="italic">lentille de eaue </hi>(
<hi rend="italic">Grant Herbier, </hi>
no269, J. Camus ds
<hi rend="smallcaps">Gdf.</hi>
<hi rend="italic">Compl.</hi>)
;
<hi rend="bold">2. </hi>
```

LENTILLE (entrée partielle)

Étymol. et Hist. 1. a) *Ca* 1170 « plante légumineuse [ici terme collectif] » (*Livre des Rois*, éd. E. R. Curtius, 2ᵉlivre, XVII, 29); **b)** xvᵉs. *lentille de eaue* (*Grant Herbier*, nº269, J. Camus ds Gdf. *Compl.*); **2.**

FIGURE 13 – Code XML – Détail de balise `<etymology>`.

5.2 La balise <etymology>

La figure 13 nous détaille le contenu de cette balise. Nous repérons dans cet exemple deux balises : <hi rend=...> et <R/>. La balise <hi rend=...>, issue du vocabulaire TEI [17], est une balise « de mise en évidence, distingue un mot ou une expression comme graphiquement distincte du texte environnant, sans en donner la raison » [18]. Elle donne donc des **informations de forme**. Avec l'attribut bold, elle spécifie un texte en gras, qui apporte néanmoins une information fonctionnelle, à savoir les différents sens du mot apparus au cours du temps. Nous allons ainsi pouvoir introduire la notion de **niveaux**, comme constaté dans la figure 11. Il existe deux autres attributs dans cet exemple : italic et smallcaps. Le premier va pouvoir être traité comme nous le verrons plus loin.

Le second est la balise <R/>, qui est utilisée lorsqu'un siècle est indiqué en chiffres romains (XVes. dans l'illustration). Il s'agit là encore d'une balise de mise en forme qui apporte indirectement une information fonctionnelle. Les siècles romains étant facilement identifiables par ailleurs, cette balise n'est pas exploitée dans l'algorithme.

Dans l'ensemble des fichiers XML étudiés, nous avons repéré au total 13 balises :

— La balise de mise en forme <hi>

— La balise indiquant un siècle <R>

— Les balises pour mettre en indice (<IND>) ou en exposant (<EXP>) du texte
Ex : `Vie du pape Grégoire</hi>, éd. H. B. Sol, A<IND>1</IND>, 599`

— Les balises pour afficher des tableaux (<table>, <tr>, <td>)

— La balise pour afficher des images <img>

— La balise <var> qui indique une variation
Ex : `ZÉPHYR, ZÉPHIR(E),<var><hi rend="italic">(ZÉPHIR, ZÉPHIRE)</hi></var>`

— La <emp> qui veut dire « emprunt »
Ex : `ZÈLE Empr. au lat. zelus ou ZÉRO <emp>Empr. à l'ital. zero`

— Les trois balises <C>, <G> et <I> qui semblent résiduelles d'un précédent encodage, et dont nous n'avons pas identifié la fonction. Au total, elles sont présentes 1, 2 et 21 fois respectivement.
Ex : `1, p. 19<C/><hi rend="smallcaps">d</hi>ds <hi rend="italic">Mittellat. W. s.v., </hi>1143, 44 : attenuatio culpae.`

Ce sont toutes des balises de mise en forme sauf les balises <emp> et <var>.

5.3 L'algorithme de création des données tabulaires

5.3.1 L'algorithme général

D'un point de vue informatique, il est important de découper le traitement en différentes procédures claires et lisibles. Il est donc préférable de faire en sorte que la longueur des codes dans chacune d'entre elles reste raisonnable. L'algorithme peut se résumer ainsi :

```
Pour chaque balise <entry> :
  Récupérer son intitulé entre les balises <form></form>
  Récupérer la catégorie grammaticale entre <gram></gram>
  Si la balise <etymology> est présente,
  pour chaque niveau, rubrique et liste des mots italiques (proc2) :
```

17. *Text Encoding Initiative*, https://tei-c.org
18. https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-hi.html

```
Identifier le siècle (proc3)
Extraire les collocations, la définition et la sous-rubrique retenue (proc4)
Séparer la partie Histoire de la partie Étymologie (proc5)
Ne garder que les collocations qui sont dans la partie Histoire
Remplir une ligne du tableau final avec :
  l'intitulé de l'entrée,
  la catégorie grammaticale,
  l'identifiant reconstruit à partir des niveaux,
  le siècle,
  la forme et ses collocations,
  la définition,
  la partie histoire.
```

L'intitulé et la catégorie grammaticale sont facilement repérables avec les balises `<form>` et `<gram>`. Leur extraction est directe. Ensuite, il s'agit de repérer la balise `<etymology>`. Si elle existe, nous recherchons si les balises `<hi rend="bold">` sont présentes pour récupérer les niveaux et les rubriques.

5.3.2 L'extraction des niveaux et des rubriques (proc 2)

L'extraction des niveaux Dans l'exemple de `lentille`, nous avons un premier niveau avec les chiffres arabes : 1, 2 et 3 et un second niveau sous le niveau 1 avec les lettres a et b.

Le premier niveau sert à mettre en évidence des différences majeures, comme l'apparition d'un nouveau sens à un siècle différent. Les sous-niveaux indiquent des différences de moindre importance à différents degrés. L'organisation des niveaux forme une structure arborescente comme l'illustre la figure 14 pour notre exemple.

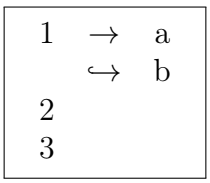


FIGURE 14 – Les niveaux de l'entrée `lentille`.

Sur l'ensemble des entrées, nous avons repéré cinq niveaux, qui respectent toujours le même ordre :

- les niveaux d'ordre 1 sont notés en chiffres romains (I, V, X) ;
- les niveaux d'ordre 2 sont notés en majuscules latines (A, B, ..., Z) ;
- les niveaux d'ordre 3 sont notés en nombres arabes (1, 2, ..., inf) ;
- les niveaux d'ordre 4 sont notés en minuscules latines (a, b, ..., z) ;
- les niveaux d'ordre 5 sont notés en minuscules grecques (α, ..., ω).

Les niveaux sont toujours à l'intérieur d'une balise `<hi rend="bold">`. Cela nous donne une première piste de traitement qui consiste à repérer les balises `<hi rend="bold">` tout en sachant que la première de ces balises contient également le texte « `Étymol. et Hist.` » et que l'identifiant d'un niveau n'est pas répété complètement. Par exemple, dans le cas de `lentille`, seul « `b)` » est contenu dans la seconde balise et non « `1. b)` ». Il sera donc nécessaire, dans un deuxième temps, de reconstituer l'identifiant complet d'une rubrique.

L'extraction du niveau est réalisé par une expression régulière qui consiste à rechercher une ou plusieurs fois :

— une lettre latine en minuscule ou
— une lettre latine en majuscule ou
— un lettre grecque en minuscule ou
— au moins un chiffre de 0 à 9 ou
— la lettre X (optionnelle) suivie de IV ou V ou VI ou VII ou VIII ou I (optionnel) ou IX

L'extraction des rubriques Une rubrique est constituée de l'ensemble du texte qui commence à la balise `<hi>` indiquant le niveau correspondant (nombre romain, majuscule latine, ...) et s'arrête juste avant la balise `<hi>` indiquant le niveau suivant.

Dans le cas de `lentille`, nous retrouvons donc 4 rubriques selon les quatre balises de mise en forme. En parallèle de ce traitement, nous récupérons tout le contenu en italique (c'est-à-dire dans une balise `<hi rend="italic">`). Nous obtenons les résultats suivants présentés dans le tableau 6.

Niveau	Rubrique	Mots en italique
1. a)	*Ca* 1170 « plante légumineuse [ici terme collectif] » (*Livre des Rois*, éd. E. R. Curtius, 2elivre, XVII, 29) ;	['Ca', 'Livre des Rois,']
b)	xves. *lentille de eaue* (*Grant Herbier*, no269, J. Camus ds Gdf. *Compl.*) ;	['lentille de eaue', 'Grant Herbier,', 'Compl.']
2.	*ca* 1300 *lantille* « tache de rousseur » (*Chirurgie de l'abbé Poutrel*, éd. Ö. Södergård, 452 ro) ;	['ca', 'lantille', 'Chirurgie de l'abbé Poutrel,']
3.	1690 opt. (Fur.). Du lat. pop. *lenticula*, class *lenticula* la « lentille (plante et graine) ; taches de rousseur » (dimin. de *lens, lentis* « lentille ») ; cf. avec 3 1637 verre en forme de lentille (Descartes, *Dioptrique*, V ds *Œuvres*, éd. F. Alquié, t. 1, p. 686).	['lenticula,', 'lenticula', 'lens, lentis', 'cf.', 'verre en forme de lentille', 'Dioptrique,', 'Œuvres,']

TABLEAU 6 – Le résultat de la procédure 2 sur l'entrée `lentille` : quatre niveaux et rubriques sont extraits, avec les mots en italique associés.

En étudiant ces rubriques dans quelques fichiers, à ce niveau de développement, nous avons constaté que :

— très souvent, elles commencent par une date ou un siècle,
— la définition apparaît entre guillemets,
— l'ensemble des mots en italique représentent la forme de l'entrée avec ses collocations en fonction du niveau.
— à la fin de la dernière rubrique, la partie étymologie n'est pas encore séparée de la partie histoire (dans la figure 6, la partie étymologie va de "Du lat. [...]" à la fin du texte.

5.3.3 L'extraction des niveaux et des rubriques (proc 3)

Dans les rubriques, les dates apparaissent sous deux formes :

— soit sous forme de 4 chiffres de 0 à 9 (ou, plus rarement, de trois chiffres pour le 9e siècle,

— soit sous forme de chiffres romains : I, II, III, IV, ... en se terminant par « es » [19].

Une lecture de la rubrique caractère par caractère est réalisée jusqu'à trouver une première suite de 4 chiffres ou des chiffres romains. Il faut ensuite vérifier que cette sous-chaîne est une date valide. Pour ce faire, un ensemble d'expressions régulières sont utilisées [20]. Le programme considère actuellement qu'une date valide est constituée soit au moins d'un chiffre de 0 à 9, soit au moins un des caractères "xviXVI" suivi de la chaîne de caractères "es".

Dans une très grande majorité des cas, l'extraction est correcte. Il reste quelques erreurs à corriger. Par exemple, certaines rubriques n'ont pas d'information de siècle, mais la procédure interprète incorrectement des valeurs numériques comme des siècles. C'est le cas de l'entrée CHIEN DE MER, dont l'unique rubrique est "`Étymol. et Hist. Cf. chien¹ étymol. B 1.`". Dans ce cas, la procédure associe le premier siècle à la rubrique.

5.3.4 Identification de la définition et des collocations d'une rubrique (proc 4)

Dans cette partie du traitement, nous avons la rubrique (appelons-la `rub`) à traiter d'une part sous forme de chaîne de caractères sans aucun formatage ni balise et d'autre part, une liste des mots ou suite de mots présents dans la rubrique sous forme italique (soit `l_italic`). Tout d'abord nous supprimons le dernier caractère des éléments de la liste `l_italic` s'il s'agit d'une virgule. Ensuite, `rub` est découpée en sous-rubriques selon le point-virgule. Nous récupérons la première sous-rubrique (appelée `sous-rub`) qui n'est pas entièrement entre crochets. Nous supprimons de la liste `l_italic` les mots qui sont dans les sous-rubriques non retenues (et non pas l'inverse c'est-à-dire supprimer de la liste `l_italic` les mots qui ne sont pas dans `sous-rub` car certains mots en italiques sont inclus dans d'autres suites de mots en italique). Nous repérons les caractères compris entre des guillemets. S'il y en a, ils sont stockés dans la variable `definition`. Nous supprimons de `sous-rub` les parties entre parenthèses. Nous effectuons le traitement des guillemets avant celui des parenthèses car certaines définitions (comme les entrées *berceau* ou *bélier*) ont des parenthèses à l'intérieur de leur définition. Nous gardons de la liste `l_italic` les mots ou suite de mots qui ne sont pas dans la liste suivante :

> '`(Lar.19`', '`FEW`', '`Compl.`', '`e`', '`i`', '`Ca`', '`ca`', '`).`', '`)`', '`(Trév.)`', '`Trév.`', '`ibid.`', '`(Ac.)`', '`Ac.`', '`la`', '`id.`', '`TLL s.v.`'

Les mots de cette liste ont été insérés dans des balises `<hi rend="italic">` mais ils ne nous apportent pas d'informations supplémentaires sur la définition de l'entrée.

5.3.5 Séparation des parties Histoire et Étymologie (proc 5)

Dans cette procédure, nous recherchons dans la rubrique la séquence (ou suite de caractères) « point suivi de deux espaces suivi d'un caractère en majuscule de A à Z ». À gauche de cette séquence, nous récupérons la partie Histoire, à droite la partie Étymologie

Pour l'exemple précédent *lentille* et les 4 rubriques détectées (cf tableau 6) : nous avons la partie Histoire égale à "`1690 opt. (Fur.)`" et la partie Etymologie à "`Du lat. pop.`

19. Apparemment les rédacteurs ont voulu gagner de la place en supprimant l'espace entre e (pour « ième » et s (pour « siècle »)

20. https://docs.python.org/fr/3/howto/regex.html

`lenticula, class. lenticula « lentille (plante et graine); taches de rousseur » (dimin. de lens, lentis « lentille »); cf. avec 3 1637 verre en forme de lentille (Descartes, Dioptrique, V ds Œuvres, éd. F. Alquié, t. 1, p. 686)."`

Après l'étude de certaines entrées pour lesquelles ce critère n'était pas suffisant, nous avons, dans un second temps ajouté une séparation des deux parties avec les schémas suivants : `1 empr.`, `1 d'apr.`,`1 dér. de`,`1 et 2 prob. empr.`,`Prob.`, `Empr.`, `Dér.`, `1 est composé de`, `Étymol.`, `Du lat.`. Ces nouveaux critères de séparations ont été ajoutés suite à l'étude du fichier XML de départ TLF-17. Un nouveau test est en cours avec les nouveaux fichiers tableurs générés, en particulier avec le fichier TLF-27.xml

6 Bilan d'étape et illustration d'un champ d'étude dérivé

6.1 La poursuite du chantier en cours

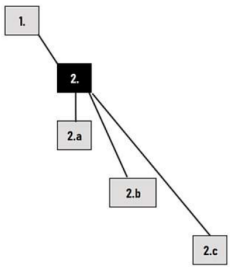
Dans l'état d'avancement actuel du projet de modélisation graphique de la polysémie évolutive des 20 000 vocables, le tableau sélectionnant et ordonnant les données fournies par les entrées H-É (fig. 15b) comporte autant de lignes que de rubriques et cinq colonnes (l'identifiant avec ses niveaux hiérarchiques, le siècle de première attestation, la forme graphique du vocable ou la collocation en cause, une définition classique entre guillemets en chevrons et l'intégralité de la rubrique).

Dans le graphe de la figure 15a, les nœuds sont disposés sur l'axe vertical de haut en bas à partir du nœud-source et sur l'axe horizontal selon la succession des siècles. Concrètement le graphe historique du n.f. BERGAMOTE comporte un nœud-source ① au 14^{e} siècle d'où dérive un nœud générique reconstitué ② au 17^{e} s., lequel a trois nœuds-fils ②a, ②b et ②c respectivement au 16^{e}, 18^{e} et 20^{e} s.

Le graphe historique de BERGAMOTE est du type illustré plus haut en figure 6. Il présente l'avantage que la disposition de chaque nœud est préétablie par le croisement des deux critères (la succession des rubriques de haut en bas et celle des siècles de gauche à droite) et que chaque place ne peut accueillir qu'un seul nœud. Le principe de sélection des arêtes est ce que Darmesteter (1887) appelait le 'rayonnement'. Dans un seul cas, celui du rattachement au nœud-source d'un premier nœud de même rang (ex. ① ⇒ ②) l'arête qui les relie illustre un enchaînement par nécessité, mais si la polysémie évolutive de BERGAMOTE comportait un ou plusieurs nœuds de même rang que le noeud-source, le rayonnement serait de nouveau privilégié : ③, ④, etc. seraient dérivés de ① et non ③ de ② et ④ de ③.

La priorité du rayonnement sur l'enchainement doit être entendue comme organisant automatiquement un graphe par défaut, mais il appartient aux historiens de la langue de fournir éventuellement des arguments pour rattacher par exemple le nœud ②.c au nœud ②.b (enchainement) plutôt qu'au nœud ②.a (rayonnement). C'est d'ailleurs manifestement le cas, car l'essence de bergamottes provenait de la variété de poires désignée en ①, alors que le bonbon désigné en ②.c dérive par métonymie (substance ⇒ objet fabriqué à partir de cette substance) de ②.b : « espèce de petit citron ».

Les lecteurs attentifs auront remarqué que le tableau a une ligne de moins que le nombre de nœuds dans le graphe. La raison en est que – dans le format actuellement testé – le nœud reconstitué ② ne figure pas dans le tableau. Deux arguments opposés plaident en faveur de l'introduction ou de l'exclusion des nœuds génériques reconstitués dans le tableau des données. L'argument favorisant leur prise en compte est le souci de



(a) Graphe

Niveaux	Siècle	Forme/colloc	Définition	Histoire
1	16		« variété de poire »	1536 « variété de poire » (Rabelais, Le Tiers Livre, éd.
2.a	17	essence de Bergamottes		1694 essence de Bergamottes (Pomet dans Les Remarques tres-curieuses.
2.b	18	bergamote,	« espèce de petit citron »	1740 bergamote « espèce de petit citron » (Ac.);
2.c	20		« bonbon parfumé à la bergamote »	1948 « bonbon parfumé à la bergamote » (Nouv. Lar. univ.).

(b) Tableau

FIGURE 15 – Modélisation graphique de la polysémie évolutive du n.f. BERGAMOTE

concordance entre le nombre des nœuds du graphe et le nombre des rubriques du tableau, que celles-ci soient effectivement présentes dans l'entrée ou reconstituées. L'argument contraire est que la recherche du noyau de sens commun aux rubriques subordonnées risque d'être vaine. Encore une fois, c'est bien le cas pour le nœud (2) puisque le fruit en cause est une variété soit de poire dans le nœud (2.a), soit de citron dans les nœuds (2b), le seul trait commun serait « variété de fruit », mais ce trait ne serait pas pertinent pour justifier un véritable regroupement entre (2.a) et (2.b)/(2.c).

6.2 Test de l'hypothèse d'une corrélation entre le volume de l'entrée H-É d'un vocable en nombre de rubriques et sa fréquence relative dans FRANTEXT au fil des siècles

Les extensions de sens accroissent la polysémie d'un vocable et même si celles-ci [1] sont partiellement composées par le déclin, voire l'extinction de certains types d'emploi et [2] certaines extensions de sens dont le sens source commun est sorti de l'usage supposent un retraitement homonymique [21], les extensions l'emportent généralement sur les extinctions et la polysémie évolutive du vocable croît de siècle en siècle. Si c'est bien le cas (et les entrées historiques du TLFi nous fournissent des indications précieuses sur cet accroissement avec la datation de la 1ère attestation de chaque nouveau type d'emploi), on peut s'attendre à ce que la fréquence relative [22] des occurrences du vocable dans les corpus séculaires s'accroisse également [Hypothèse 1] et aussi que l'augmentation de cette fréquence relative soit plus forte quand plusieurs extensions de sens sont attestées durant un même siècle et plus faible quand une seule extension de sens, voire aucune, n'est attestée durant un siècle [Hypothèse 2]. Pour tester ces deux hypothèses nous avons sélectionné dix vocables dont la polysémie s'est particulièrement accrue au fil des siècles :

— 3 noms : *carte, jeu, jour*
— 4 adjectifs : *clair, juste, propre, sec* [23]
— 3 verbes : *casser, jeter, monter*

Pour chaque vocable nous avons comparé le nombre cumulé des rubriques de leur entrée historique siècle par siècle (NbR+) [24] et la fréquence relative du vocable dans chaque

21. Voir l'analyse historique de la polysémie du n.m. *timbre* par Darmesteter (1887 : 81–83) qui n'est plus plausible au 21e siècle, car certains emplois comme le *timbre* d'un casque sont sortis de l'usage courant (à l'exception des spécialistes de l'histoire militaire).

22. C'est-à-dire le nombre des occurrences du vocable de siècle en siècle divisé par le volume de chaque corpus séculaire.

23. affublés d'emplois substantivés, ex. tirer les choses au CLAIR ; le JUSTE et l'injuste / réécrire au PROPRE ; être à SEC / fumer une SÈCHE

24. Il est à noter que pour associer à chaque siècle le nombre de rubriques en cause, il faut prendre deux précautions : [1] chaque rubrique représente un sens attesté à partir d'un siècle et éventuellement des

corpus séculaire de sa première attestation jusqu'au 20e siècle (PrScl). Pour que ces deux courbes figurent aisément sur un même graphique, la fréquence relative a été multipliée par 100. Chaque courbe est accompagnée d'une « courbe de tendance », ici en l'occurrence une droite. Trois classes comparatives se dégagent en fonction de l'orientation concordante *vs* discordante des deux droites de tendance.

La droite de tendance du nombre cumulé de rubriques de l'entrée H-É du TLFi a nécessairement une orientation croissante : ↗. Si celle de la présence du vocable dans FRANTEXT de siècle en siècle est du même type, le profil couplé du vocable est de la classe I à condition que sa présence séculaire n'ait pas été perturbée par une ou plusieurs fluctuations, sinon il est de la classe II. En revanche, si la droite de tendance de la présence séculaire du vocable a une orientation décroissante : ↘, les deux tendances sont discordantes (classe III) et il y a lieu de rechercher des facteurs supplémentaires qui ont pu causer cette décroissance.

Classe I Concordance entre la droite de tendance du nombre cumulé des rubriques et celle de la présence séculaire du vocable dans FRANTEXT sans fluctuation ⇒ *carte, casser, jeu.*

Pour les deux noms *carte* (figure 16) et *jeu* (figure 17) et pour le verbe *casser* (figure 18), l'orientation des deux tendances est apparentée : ↗↗ et aucune fluctuation n'a perturbé la tendance croissante de la présence séculaire des trois vocables.

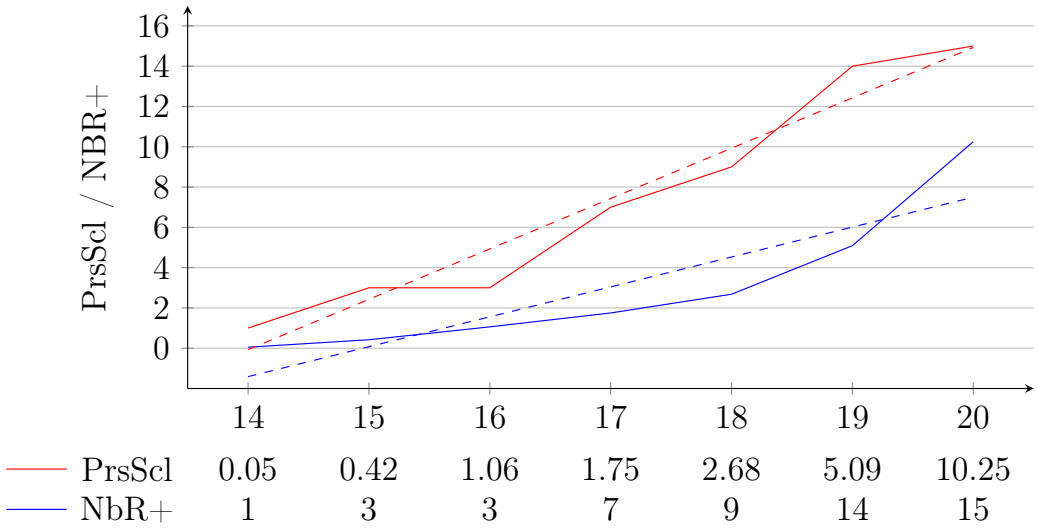


FIGURE 16 – Profil historique couplé du n.f. CARTE

Classe II Concordance entre la droite de tendance du nombre cumulé des rubriques et celle de la présence séculaire du vocable dans FRANTEXT au prix de fluctuations ⇒ *jeter, jour, juste, propre, sec*

La classe II comporte deux vocables de plus que la classe I, avec le nom *jour*, les adjectifs *juste, propre* et *sec*, et le verbe *jeter*. Comme précédemment les tendances des

spécications de ce sens les sens survenues ultérieurement, nous ne tenons donc compte que de la première attestation de ce sens, et [2] il en est de même pour les sous-rubriques, si bien que la consultation linéaire des entrées H-É implique des retours en arrière. On peut s'en convaincre en examinant le tableau **??** où le sens classé en 1.b) lentille de eaue est attesté au 15e siècle, alors que le sens classé en 2 (lantille « tache de rousseur ») est attesté vers 1300. Seul le tableau complet de toutes les premières attestations, indépendamment de leur classement dans l'entrée H-É permet donc d'associer un nombre pertinent de sens à chaque siècle

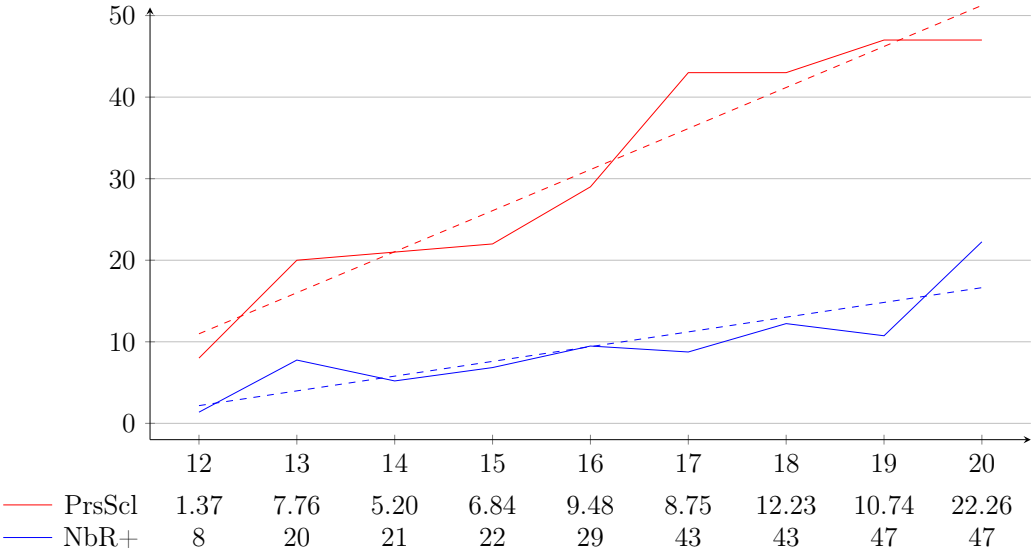


FIGURE 17 – Profil historique couplé du n.m. JEU

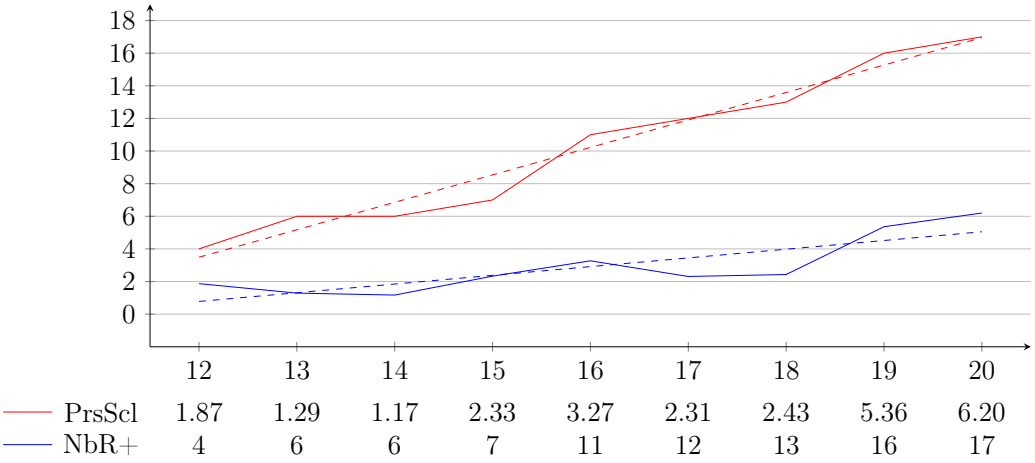


FIGURE 18 – Profil historique couplé du v. CASSER

deux courbes sont apparentées : ↗↗, mais on constate des fluctuations de la présence séculaire du vocable.

Pour *jeter* (figure 19), ce sont la croissance subite de 15,73 à 35,18 entre le 15e et le 16e siècle, puis inversement la décroissance un peu moins marquée de 41,83 à 26,08 entre le 19e et le 20e siècle.

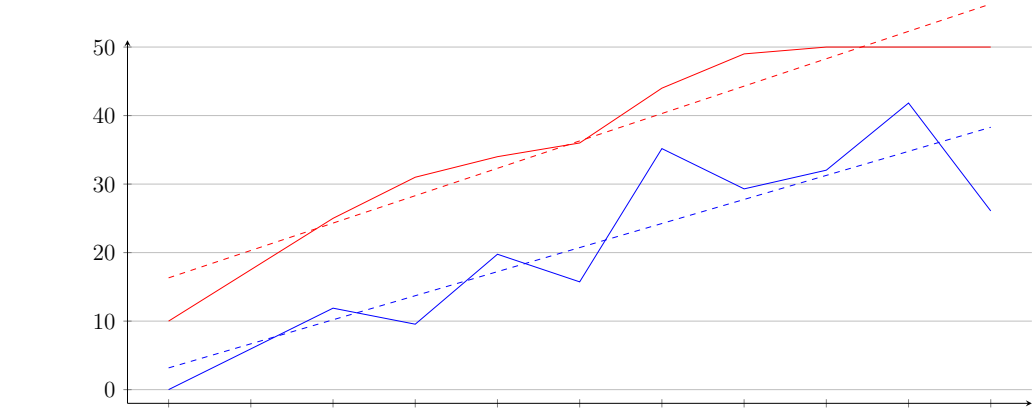


FIGURE 19 – Profil historique couplé du v. JETER

Pour *jour* (figure 20), c'est la forte croissance étalée sur deux siècles, entre le 13e (7,66) et le 15e (27,09), et inversement la forte décroissante entre le 15e (27,09) et le 16e siècle (14,38).

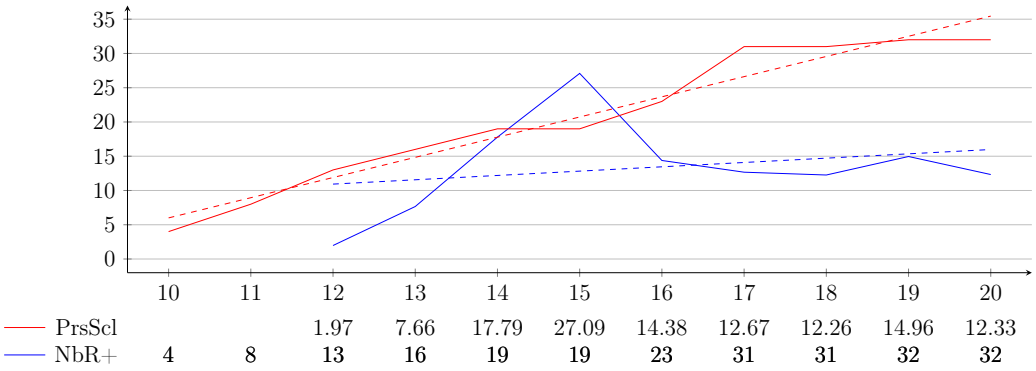


FIGURE 20 – Profil historique couplé du n.m. JOUR

La courbe de présence séculaire de *juste* (figure 21) présente des fluctuations encore plus importantes dans le sens de la croissance entre le 13e (2,07) et le 14e siècle (27,25), puis entre le 16e (14,16) et le 17e siècle (29,97) avant une forte décroissance étalée entre le 17e (29,97) et le 19e siècle (15,08).

La courbe de présence séculaire de l'adj. et n.m. *propre* (figure 22) est caractérisée par une période de trois siècles de stabilisation à un haut niveau (entre 38,15 et 41,76) après une brusque croissance (de 16,64 à 38,15 entre le 15e et le 16e siècle) et avant une aussi brusque décroissance entre le 18e et le 19e (de 41,76 à 23,15).

Enfin la courbe de présence séculaire de *sec* (figure 23) est marquée par un double zigzag entre des valeurs basses au 13e (2,37), 15e (2,30) et 17e siècle (4,55) et deux valeurs hautes au 14e (17,42) et au 16e siècle (19,09).

L'examen détaillé des causes possibles de ces fluctuations est prématuré, l'inventaire des facteurs imaginables étant élevé. Par contre, cet examen s'impose pour les deux vo-

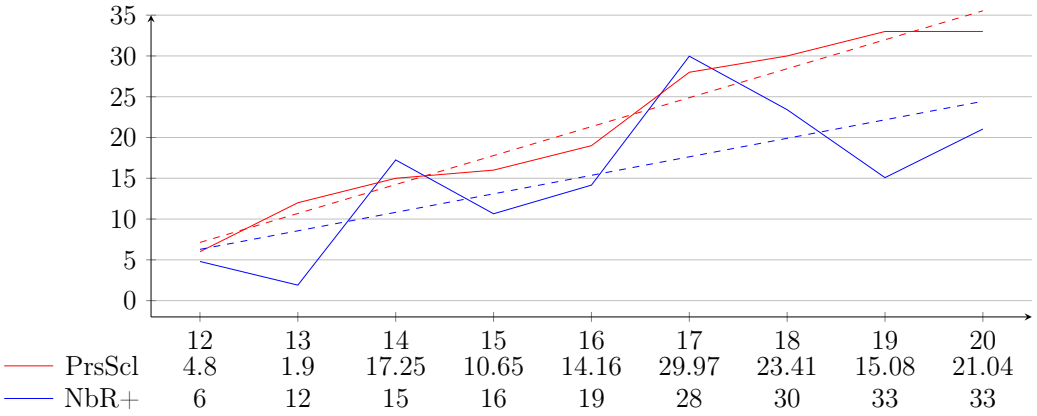


FIGURE 21 – Profil historique couplé de l'adj. et n.m. JUSTE

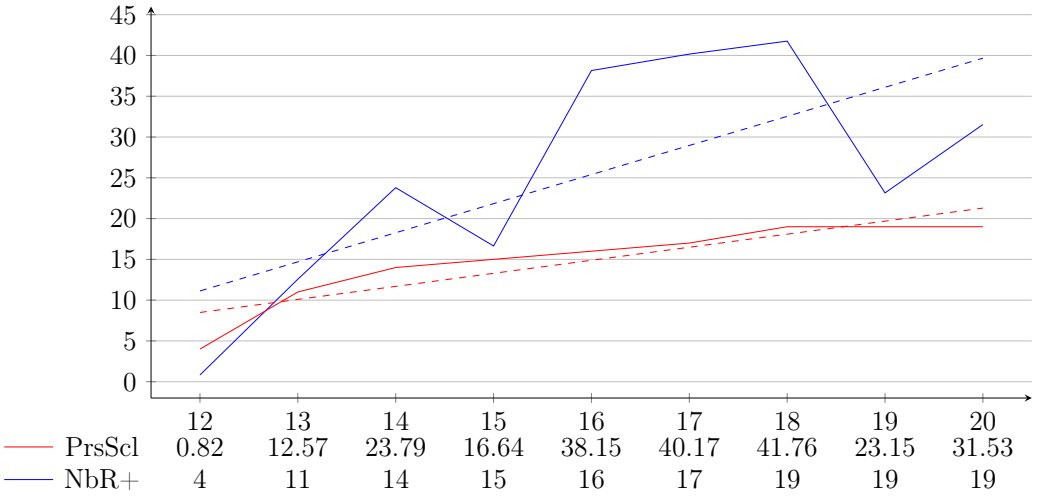


FIGURE 22 – Profil historique couplé de l'adj. et n. PROPRE

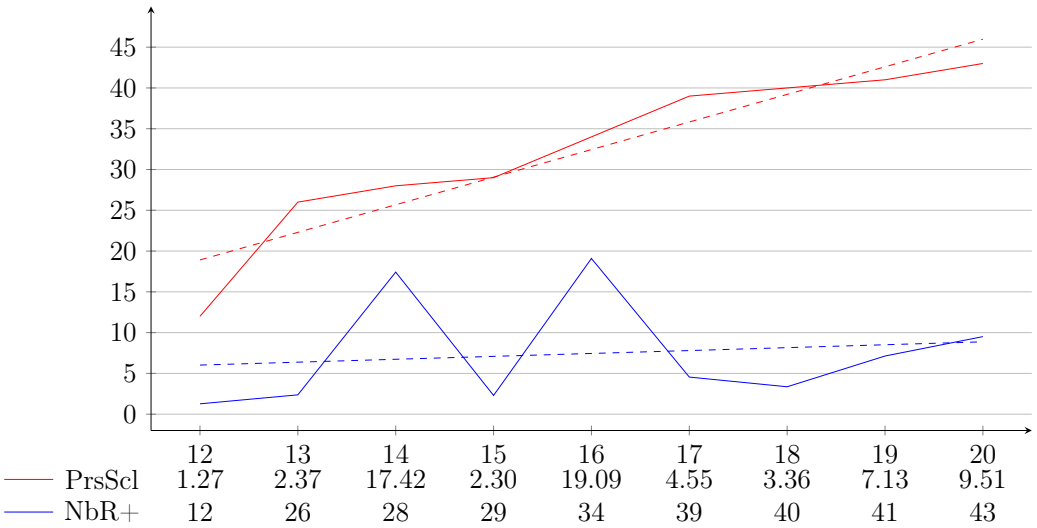


FIGURE 23 – Profil historique couplé de l'adj. SEC

cables qui relèvent de la classe III, puisque dans ces deux cas, la tendance des deux courbes est discordante : ↗↘.

Classe III Discordance entre la droite de tendance du nombre cumulé des rubriques (progression) et celle de la présence séculaire du vocable dans FRANTEXT (régression).

Les valeurs de présence séculaire de l'adj., n.m. et adv. *clair* (figure 24) très élevées pour le 12e s. (32,66) et surtout le 13e s. (46,44) peuvent sans doute s'expliquer par l'ambivalence de la forme ***clers*** qui désigne fréquemment un clerc (<*lat. clericus*) au lieu de l'adj. *clair* (<*clarus*). Cette ambivalence vaut jusqu'au 13e s., ce qui pourrait expliquer qu'à partir de sa disparition au 14e siècle, on observe une chute drastique de la présence séculaire (22,30). Il est vrai que la croissance de la courbe du nombre cumulé des rubriques est très faible à partir du 13e siècle (6 nouvelles rubriques sur 8 siècles, soit moins d'une extension de sens par siècle), ce qui laisserait entendre que la faible proportion de rubriques nouvelles aurait un effet supérieur à celui supposé dans l'hypothèse 2, ne se limitant pas à réduire la tendance croissante, mais allant jusqu'à l'inverser.

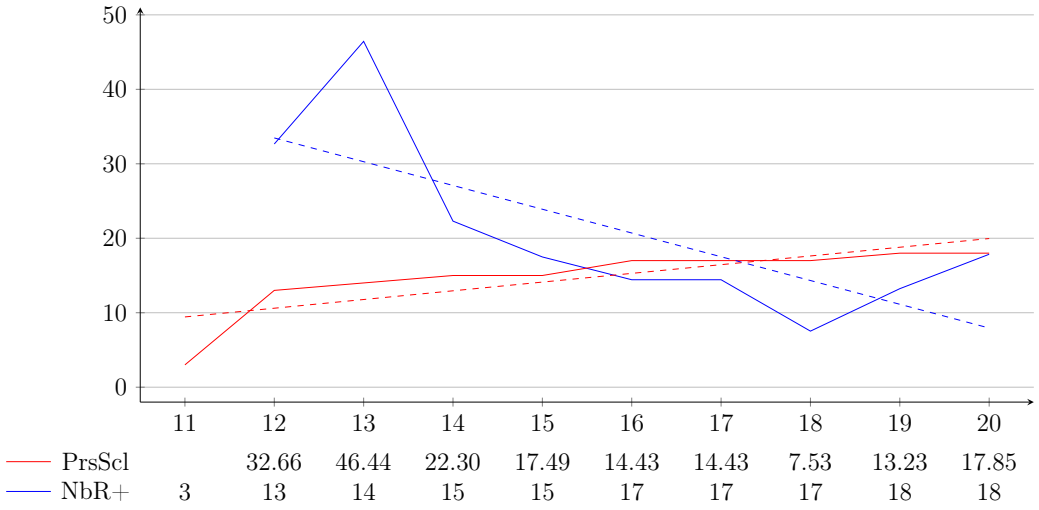


FIGURE 24 – Profil historique couplé de l'adj. et n.m. CLAIR

Un facteur se laisse facilement tester, celui de la montée en puissance d'un synonyme (contextuel). La figure 25 représente le rapport entre la fréquence de *visible* et celle de *clair* entre le 14e et le 21e siècle. On constate effectivement un accroissement de ce rapport de 4 à 17% entre le 15e et le 17e siècle qui pourrait compléter la levée de l'homographie clers pour l'adj. *clair* et le substantif *clerc* à partir du 14e siècle. Aux 19e et 20e siècles la présence séculaire de *clair* remonte au même niveau qu'au 15e siècle.

On ne peut donc pas exclure que les deux principaux facteurs du déclin de la présence de *clair* entre le 14e et le 17e siècle soient :

a) entre le 13e et le 14e siècle, la disparition de l'homographie entre CLERS_1 (le clerc) et CLERS_2 (l'adj. *clair*)

b) et entre le 15e et le 18e s. la concurrence entre les adj. CLAIR et VISIBLE (TRANSPARENT et PERCEPTIBLE ayant toujours eu des fréquences trop faibles pour concurrencer efficacement *clair*.

Cependant seule une analyse comparative minutieuse des contextes dans lesquels *clair* et *visible* ont figuré entre le 15e et le 18e siècle pourrait fournir des arguments plausibles en faveur de cette hypothèse.

De son côté, la courbe de tendance séculaire du v. *monter* a une tendance décroissante et en outre elle est très fluctuante (cf. fig. 26 et François 1980, 2010). Un seul changement

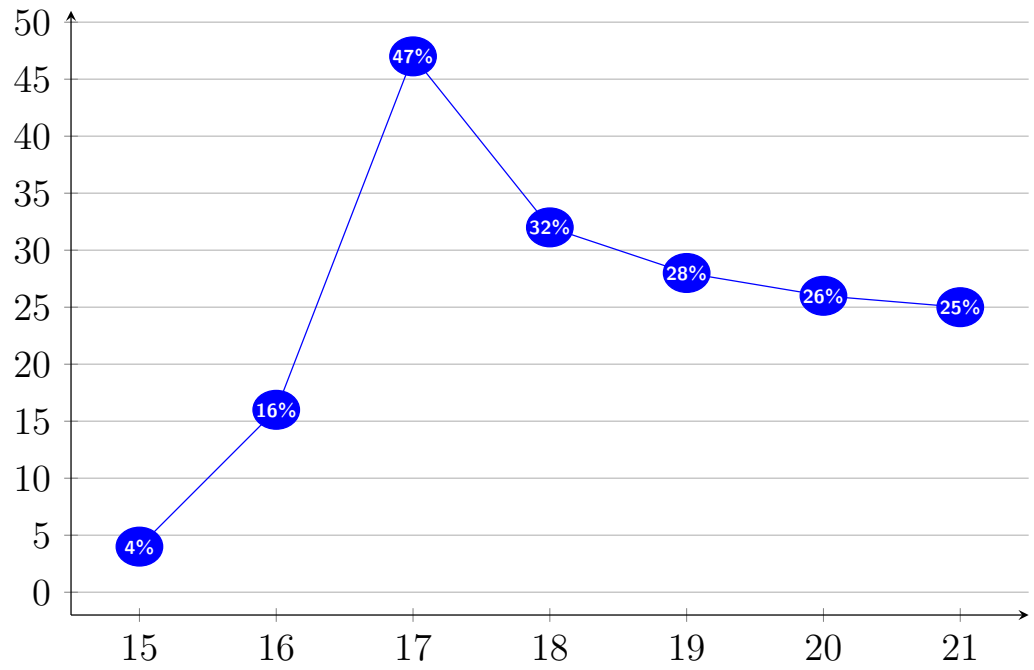


FIGURE 25 – Rapport entre les deux fréquences de VISIBLE et de CLAIR entre le 15e et le 21e siècles

sociétal remarquable vient à l'esprit, le déclin de l'usage du cheval comme monture, très important à l'époque de la chevalerie et de la fonction militaire de la cavalerie. La pratique de l'équitation ne suffit sans doute pas à compenser cette fonction beaucoup plus prégnante par le passé.

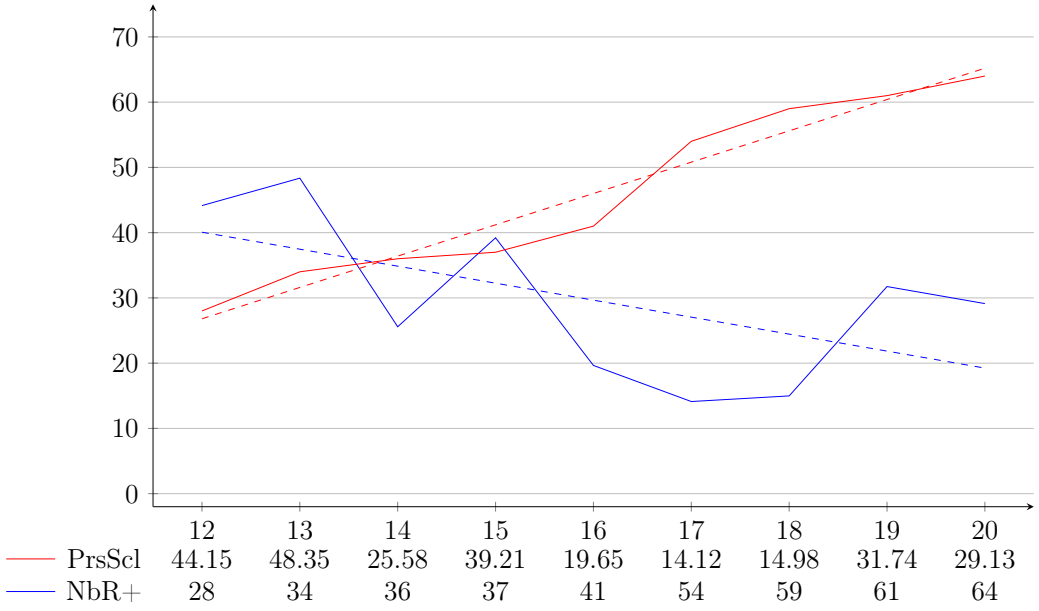


FIGURE 26 – Profil historique couplé du v. MONTER

Pour tester l'hypothèse du déclin de l'usage du cheval comme monture ou comme bête de trait, nous avons eu recours à la fonction de repérage des cooccurrences de FRANTEXT et nous avons calculé pour chaque siècle le rapport entre les occurrences du v. *monter* quel que soit le contexte et les cooccurrences entre toute forme du verbe et les formes *cheval/chevaux* à une distance comprise entre 1 et 3 mots dans le contexte droit. Le résultat est étonnamment parlant (fig. 27).

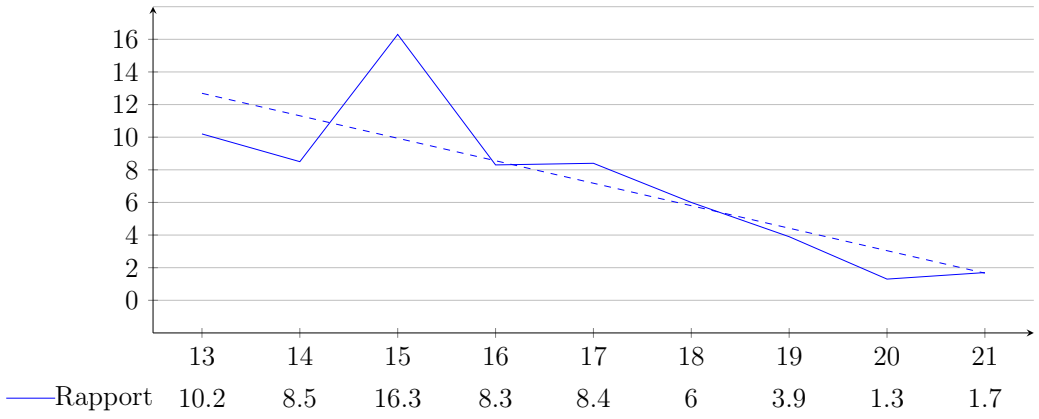


FIGURE 27 – Rapport entre les deux fréquences de MONTER et MONTER [...max.3 mots dont cheval/chevaux...] entre le 13e et le 21e siècles

Si l'on compare l'orientation de la droite de tendance dans les deux figures 26 et 27, on constate qu'elles sont étroitement apparentées. On ne peut pas exclure que d'autres facteurs secondaires aient contribué à l'orientation décroissante de la présence séculaire du v. éemonter, mais il est très vraisemblable que le déclin progressif de la fonction sociétale du cheval dans la société française ait joué un rôle décisif. Il resterait toutefois à élucider le pic de cooccurrences entre *monter* et *cheval* au 15e siècle !

Au final, il apparaît donc que la base de données textuelles FRANTEXT fournit des outils d'analyse diachronique quantitative qui permettent de tester l'hypothèse d'une corrélation entre la fréquence relative d'un vocable au fil des siècles (désignée comme sa « présence séculaire ») et l'accumulation des rubriques de l'entrée H-É de ce vocable dans le TLFi. Et globalement, sur les dix vocables sélectionnés, trois tendent à valider cette hypothèse, cinq s'en rapprochent au prix de fluctuations que d'autres facteurs secondaires pourraient éclairer, et deux présentent une discordance avec l'hypothèse que des facteurs secondaires (un pour MONTER, deux pour CLAIR) parviennent à éclairer. Cependant, il va de soi que la validation de cette hypothèse ne peut pas être assurée pour au moins deux raisons :

1. les entrées H-É du TLFi ne mentionnent pas de date de déclin du sens attesté en premier et/ou des extensions de sens, et il est probable que, ici ou là, l'extinction d'un sens a eu un effet sur la présence séculaire du vocable examiné ;
2. l'importance sociétale d'une extension de sens d'un vocable ne peut pas ressortir de son entrée H-É, son évaluation suppose un examen approfondi des textes enregistrés et notamment des collocations dans lesquelles ce vocable figure.

Il s'agit là des limites d'une analyse sémantaxique quantitative : elle ouvre des portes, mais il reste à explorer ce que cachaient ces portes.

Bibliographie

Études lexicologiques

BERNARD Pascale « Le Trésor de la langue française informatisé », *Traduire pour le théâtre* (Hors cahier thématique), Traduire n° 222 : 125-136

BERNARD Pascale / DENDIEN Jacques / PIERREL Jean-Marie (2004), « A Computerized Dictionary : Le trésor de la langue française informatisé (TLFi) ». In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pages 40–43, Geneva, Switzerland. COLING https://aclanthology.org/W04-2107.pdf.

BLANK Andreas (1997), *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen.* [Principes du changement des sens lexicaux illustrés par les langues romanes] Berlin : De Gruyter [Beihefte zur Zeitschrift für romanische Philologie].

BLANK Andreas (2000), « Pour une approche cognitive du changement sémantique : aspect sémasiologique ». In J. François (dir.), *Théories contemporaines du changement sémantique*. Mémoire IX de la Société de Linguistique de Paris : 59-74. https://bibliographie.uni-tuebingen.de/xmlui/handle/10900/78083

CARON Philippe / DEFIOLLE Rodolphe / LAY Marie-Hélène (2019) *L'enjeu des métadonnées dans les corpus textuels ; un défi pour les sciences humaines*, Rennes, P.U.R, Collections Rivages linguistiques

DARMESTETER Antoine (1887) *La Vie des mots étudiée dans leurs significations*, Paris, Delagrave [rééd. Paris, Champ Libre, 1979] https://catalogue.bnf.fr/ark:/12148/cb31995071n

FRANÇOIS Alexandre (2008), "Semantic maps and the typology of colexification. Intertwining polysemous networks across languages", in : Vanhove, Martine (ed.), *From polysemy to semantic change. Towards a typology of lexical semantic associations*, Amsterdam / Philadelphia, Benjamins : 163–216.

FRANÇOIS Alexandre (2022), "Lexical tectonics. Mapping structural change in patterns of lexification", in : Georgakopoulos, Thanasis/Polis, Stéphane (eds.), *The future of mapping. New avenues for semantic maps research*, Zeitschrift für Sprachwissenschaft 41(1) :89–123

FRANÇOIS Jacques (1980), « Le lexique verbal français et les dégroupements homonymiques ». *Zeitschrift für französische Sprache und Literatur*, XC-(1) : 1-24.

FRANÇOIS Jacques (2005), « Quand jouer, c'est jouer de la musique : Repérage contextuel de quatre zones de l'espace sémantique du verbe JOUER ». In : E. Richard, M.C. Le Bot, M. Schuwer F. Neveu (dir.) *Aux marges des grammaires – Mélanges en l'honneur de Michèle Noailly*. Rennes : P.U.R. : 143-158.

FRANÇOIS Jacques (2008), *Une approche diachronique quantitative de la polysémie verbale*. Cahier du CRISCO 24. Université de Caen-Normandie. https://hal.archives-ouvertes.fr/hal-01870490v1

FRANÇOIS Jacques (2009), "L'évolution de la polysémie verbale documentée à partir des corpus textuels et des exemples lexicographiques". Travaux linguistiques du CerLiCO 23, *L'exemple et le corpus - Quel statut ?*, 181-200.

FRANÇOIS Jacques (2010a), « L'évolution de la polysémie verbale documentée à partir des corpus textuels et des exemples lexicographiques ». *Travaux linguistiques du CerLiCO* 23 : 181-200.

FRANÇOIS Jacques (2010b), « L'étude de la polysémie verbale entre dérivation et invariance ». *Actes du 2e Congrès Mondial de Linguistique Française*. La Nouvelle-Orléans, juillet 2010 https://www.linguistiquefrancaise.org/articles/cmlf/abs/2010/01/cmlf2010_000268/cmlf2010_000268.html

FRANÇOIS Jacques (2020a) « Pour un retraitement informatisé et dynamique des notices historiques du TLFi ». *Cahiers de Lexicologie* 117 (2021) : 11–48

FRANÇOIS Jacques (2020b), « Les fluctuations historiques de la polysémie lexicale ». *Travaux de linguistique* 81 : 57-98.

FRANÇOIS Jacques (2021) « Comment visualiser l'évolution historique des polysémies lexicales : l'itinéraire sémantique de *terre* et *monde* ». *Zeitschrift für romanische Philologie* 137(3) : 625-665.

FRANÇOIS Jacques (2022), *Les techniques de visualisation dans les sciences du langage*. Manuscrit CRISCO, Université de Caen-Normandie https://normandie-univ.hal.science/hal-03797302

FRANÇOIS Jacques (à paraître) « Les extensions de sens », in Eva BUCHI (dir.), *Manuel d'Étymologie Romane*. Berlin : De Gruyter.

KOCH Peter (2000), « Pour une approche cognitive du changement lexical : aspect onomasiologique », Société de linguistique de Paris, *Mémoire* 9 : 75–95.

LEVELT Willem (1989) ; *Speaking : From intention to articulation*. Cambridge (MA) : MIT Press.

MARTIN Robert (2001), *Sémantique et automate. L'apport du dictionnaire informatisé*. Presses Universitaires de France.

MEILLET Antoine ([2]1921|[1]1906), « Comment les mots changent de sens, Année sociologique » 1906 (repris dans *Linguistique historique et linguistique générale*, Paris, Champion, 1921, 230–271)

MEL'ČUK Igor (2023), *General phraseology*. Amsterdam / Philhadelphy : Bernjamins.

PAUL, Hermann (1880), *Prinzipien der Sprachgeschichte*, Halle, Niemeyer

STERN Gustav (1931), *Meaning and change of meaning*, Bloomington :Indiana University Press

ULLMAN Stephen (1952), *Précis de sémantique française*. Presses Universitaires de France.

ULLMAN Stephen (1962), *Semantics – An introduction to the science of meaning*, Oxford, Blackwell

WIERZBICKA Anna (1996), *Semantics : Primes and universals*. Oxford / New-York : Oxford University Press.

Dictionnaires et bases de données textuelles

ACADÉMIE FRANÇAISE (1694), *Dictionnaire de l'Académie Françoise*, 2vol., Paris, J.-B. Coignard, 1694. Éditions consultables sur le site *Dictionnaire d'autrefois* de l'université de Chicago https://artflsrv04.uchicago.edu/philologic4.7/publicdicos/ : 1ère (1694), 2ème (1762), 3ème (1798), 8ème (1935).

BESCHERELLE Louis-Nicolas (1846), *Dictionnaire national ou Dictionnaire universel de la langue française*, Paris : Simon.

FRANTEXT : Base de données textuelles. ATILF(CNRS) – Université de Lorraine https://www.frantext.fr

FURETIÈRE Antoine (1690), *Dictionnaire universel* https://gallica.bnf.fr/ark:/12148/bpt6k50614b

GODEFROY Frédéric (1880-1895) *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IXe au XVe siècle*. Paris : Vieweg

HATZFELD Antoine / DARMESTETER Antoine (1890), *Dictionnaire général de la langue française*. Paris : Delagrave. https://gallica.bnf.fr/ark:/12148/bpt6k206410-m.image

LAFAYE Pierre-Benjamin (1858), *Dictionnaire des synonymes de la langue française*, Paris : Hachette [rééd. 1996, Booking International]

LITTRÉ Émile (1872), *Dictionnaire de la langue française*. Paris : Hachette.

MARTIN Robert / BAZIN Sylvie (ed.), *Dictionnaire du Moyen Français* (DMF 2015), Nancy, ATILF, 2015, http://www.atilf.fr/dmf.

NICOT Jean (1606), *Thresor de la langue francoyse, tant ancienne que moderne*. Paris : David Douceur.

REY Alain (1992), *Dictionnaire historique de la langue française*. Paris : Éditions Le Robert.

TLF(i) = IMBS Paul / Quemada Bernard (dir.), *Trésor de la langue française. Dictionnaire de la langue du XIXe et du XXe siècle (1789–1960)*, 16vol., Paris, Éditions du CNRS/Gallimard, 1971–1994,http://atilf.atilf.fr

WARTBURG, Walther von et al. (1922-2002), *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes*, [Dictionnaire étymologique français. Présentation du vocabulaire gallo-roman], 25vol., Bonn/Heidelberg/Leipzig-Berlin/Bâle, Klopp/Winter/Teubner/Zbinden, 1922–2002.