



# **La polysémie évolutive des vocables du français contemporain visualisée sur la base des entrées historiques du TLFi**

Laurette Chardon, Jacques François, Justine Reynaud

## **► To cite this version:**

Laurette Chardon, Jacques François, Justine Reynaud. La polysémie évolutive des vocables du français contemporain visualisée sur la base des entrées historiques du TLFi. Bulletin de la Société de Linguistique de Paris, 2023, 118 (1), pp.75-124. <10.2143/BSL.118.1.3292780>. <hal-04598504>

**HAL Id: hal-04598504**

**<https://hal.science/hal-04598504v1>**

Submitted on 3 Jun 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC-SA 4.0 - Attribution - Non-commercial use - ShareAlike - International License

Jacques FRANÇOIS, Laurette CHARDON, Justine REYNAUD

## **La polysémie évolutive des vocables du français contemporain visualisée sur la base des entrées historiques du TLFi**

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<sup>1</sup> 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<sup>2</sup> 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 20<sup>e</sup> 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<sup>3</sup> 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<sup>4</sup>, 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

---

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 signifiants 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/vocab>>

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 XXI<sup>e</sup> siècle.

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 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. Concrètement, 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<sup>5</sup> 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<sup>6</sup> 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  $\pm 20\,000$  entrées H-É décrivant une polysémie évolutive (§3), l'éventail typographique de ces entrées et leur interprétation fonctionnelle (§4) 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. Un aperçu des techniques de l'informatisation fonctionnelle des entrées H-É est fourni en Annexe.

## **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.

---

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, Defiollle et Lay (dir. 2019).

### 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 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, mais qui présentent un degré variable de superposition.

Le premier sémanticien à avoir proposé une représentation visuelle de la superposition entre la couverture sémantique de différents vocables est sans doute Pierre-Benjamin LAFAYE dans la Préface à son *Dictionnaire des synonymes* (1858 : xxxix).

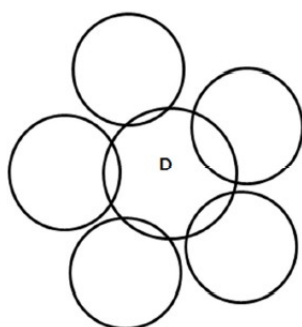


Figure 7 de Lafaye (1858) : idées communes et traits particuliers d'une famille de vocables de sens voisins. Le cercle central D, entrant en intersection avec les cinq autres, représente le sens commun aux cinq vocables, ce qui assure leur ressemblance. Si ce sens est véhiculé par le n.m. *malheur*,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

Figure 1 : Une visualisation sémantique de Pierre-Benjamin Lafaye et son interprétation

Ce mode de visualisation<sup>7</sup> 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) 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 François (2005 : 149, cf. figure 2 ci-dessous).

7 Cf. François 2022, « Les techniques de visualisation dans les sciences du langage ».

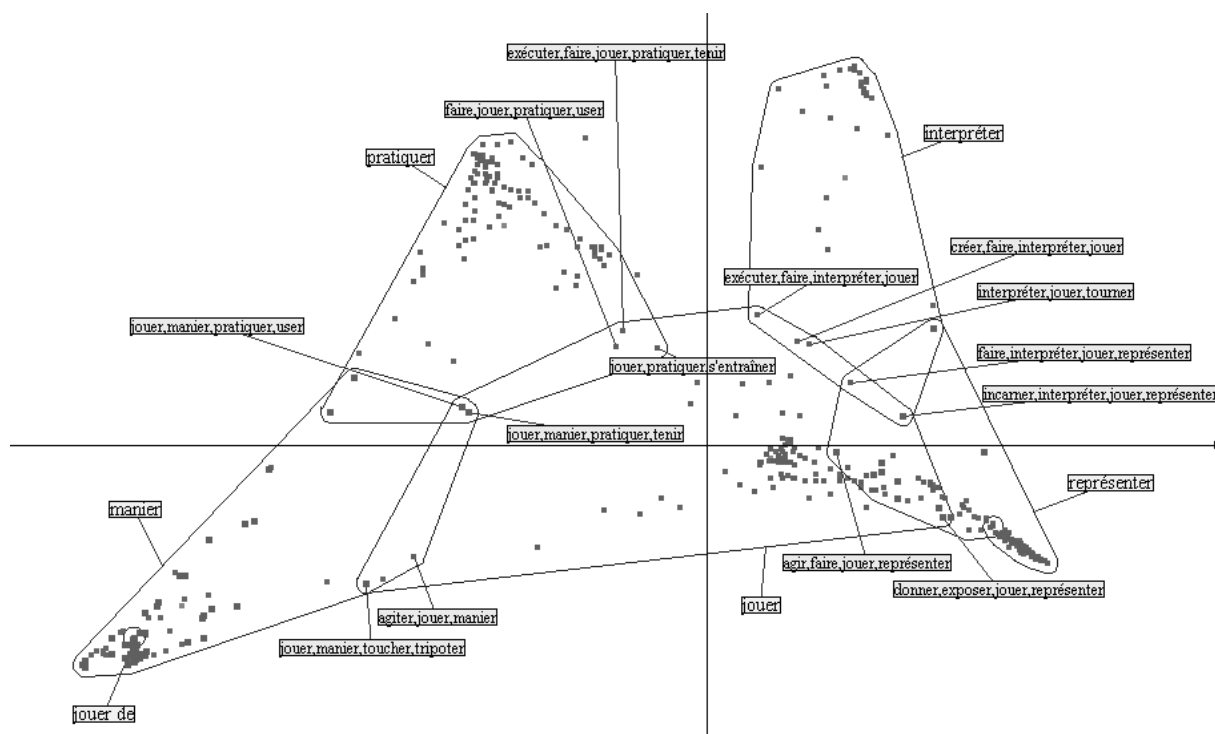


Figure 2 : Visualisation de l'intersection entre la polysémie du v. *jouer* et de quatre de ses synonymes contextuels {*interpréter*, *manier*, *pratiquer*, *représenter*} à l'aide du DICTIONNAIRE ÉLECTRONIQUE DES SYNONYMES du CRISCO

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 1 (Lafaye, 1858). Son « enveloppe convexe » est entourée de 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 appelées des « cliques de synonymes ». Ainsi, on voit que *jouer* et *manier* partagent cinq 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 2<sup>ème</sup> synonyme, *pratiquer*. De même les deux cliques 5-6 sont partagées par le 3<sup>ème</sup> synonyme, *interpréter*, et le 4<sup>ème</sup>, *représenter*.

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

En revanche les deux synonymes figurant à gauche dans la fig. 2 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. 3 ci-dessous est une version simplifiée de la fig. 2 destinée à illustrer le contexte sémantique associé aux vocables membres de chacun des deux ensembles et l'appartenance de *jouer* aux deux ensembles.

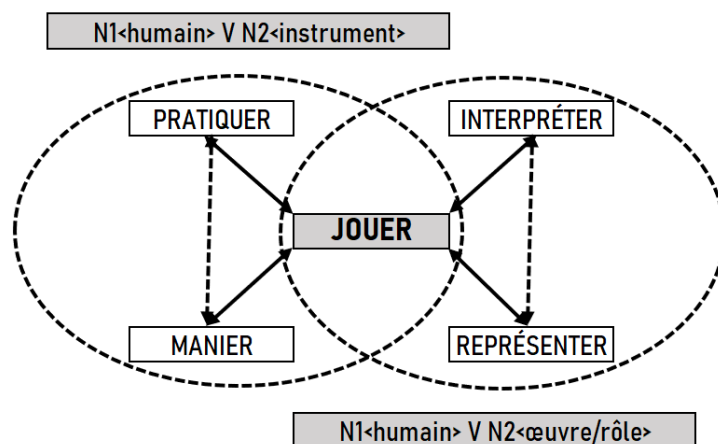


Figure 3 : Les deux ensembles de synonymes {jouer, manier, pratiquer}  
et {jouer, interpréter, représenter}

Passons maintenant à l'APPROCHE DIACHRONIQUE de la sémantique lexicale (cf. Blank 2000). La figure 4 reproduit un jeu de visualisations de trois types de polysémie évolutive emprunté à François (2008). Soit une unité lexicale L véhiculant à  $t^a$  un seul type de sens  $\langle s_1 \rangle$ . À partir de  $t^d$  (époque de diversification sémantique) L peut véhiculer, selon le contexte, un second sens  $\langle s_2 \rangle$ . 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  $t^o$ .

- Dans le premier cas, les deux types d'emploi sont toujours valides, c'est l'illustration d'une polysémie effective à  $t^o$ . L'ordre dans lequel les deux sens  $\langle s_1, s_2 \rangle$  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'utilisateur de la langue à l'époque  $t^o$  ignore généralement que L avait originellement un sens différent et est passé par une phase de polysémie avant que  $\langle s_2 \rangle$  se substitue finalement à  $\langle s_1 \rangle$ . 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  $\langle s_2 \rangle$  a décliné et a disparu, tandis que  $\langle s_1 \rangle$  s'est maintenu jusqu'à  $t^o$ . 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  $\langle s_2 \rangle$  pendant une période plus ou moins étendue (voir plus bas le sens 6 du n.m. *timbre* selon Darmesteter).

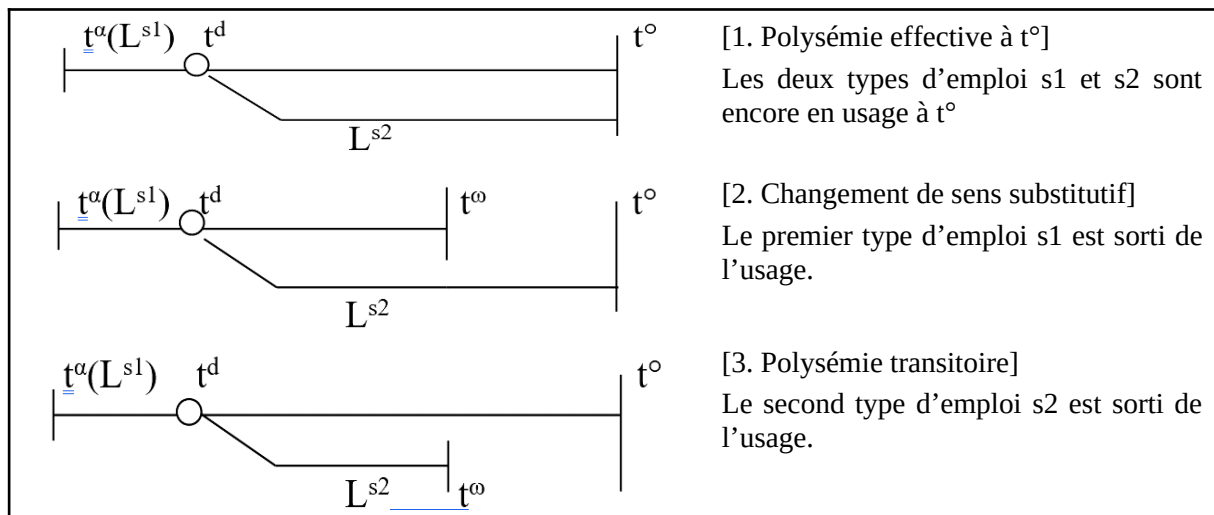


Figure 4 : Trois types de polysémie évolutive (cf. François 2008 : 10-12)

En réalité, la question est plus complexe, car il faut tenir compte des technoclectes, c'est-à-dire 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 douze 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<sup>e</sup> 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, deux techniques de visualisation ont été testées<sup>8</sup>. La première (Fig .5) subdivise l'évolution de la polysémie d'un vocable en « instantanés séculaires ». L'évolution de cette polysémie « abstraite » – c'est-à-dire sans réalisation attestée – passe par trois périodes<sup>9</sup> :

- le 12<sup>e</sup> siècle, où le sens premier général [1] apparaît ainsi qu'une dérivation [1.1] et une extension [B]
- le 15<sup>e</sup> siècle, où une seconde extension [C] émerge, tandis qu'une seconde dérivation du premier sens [1.2] s'ajoute à la première, et que le sens [B] donne lieu à deux dérivations simultanées (à l'échelle d'un siècle) [B.1/2],

<sup>8</sup> Le mode de visualisation de la polysémie évolutive illustré par la fig.5 a été appliqué à l'adj. et n.m. *curieux* (François 2020 a: 79-84 ), au n.f. *terre* (François (2020b : 27, 29), et au n.f. *batterie* (François 2021 : 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).

<sup>9</sup> 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.

- et le 19<sup>e</sup> siècle, où trois dérivations simultanées du sens [C] complètent le processus.

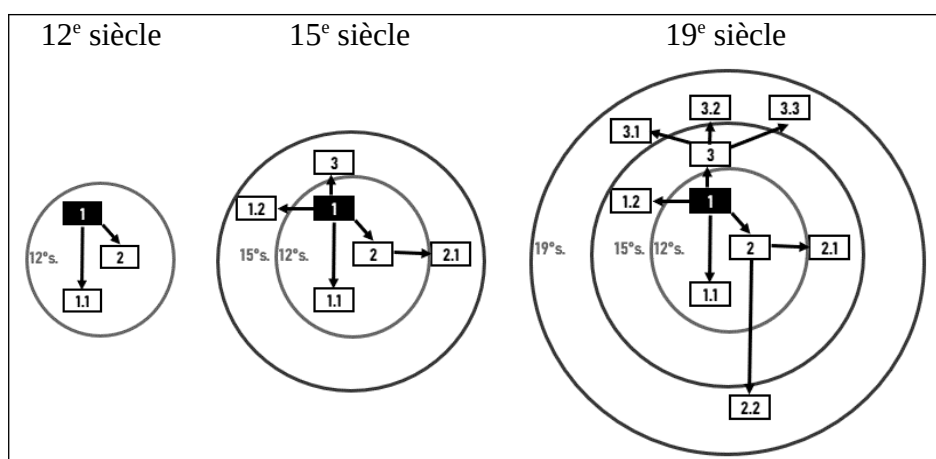


Figure 5 : Visualisation d'une polysémie évolutive abstraite par instantanés séculaires

La seconde technique consiste à construire un graphe historique dont chacun des nœuds symbolise l'un des sens répertoriés dans la fig. 5 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).

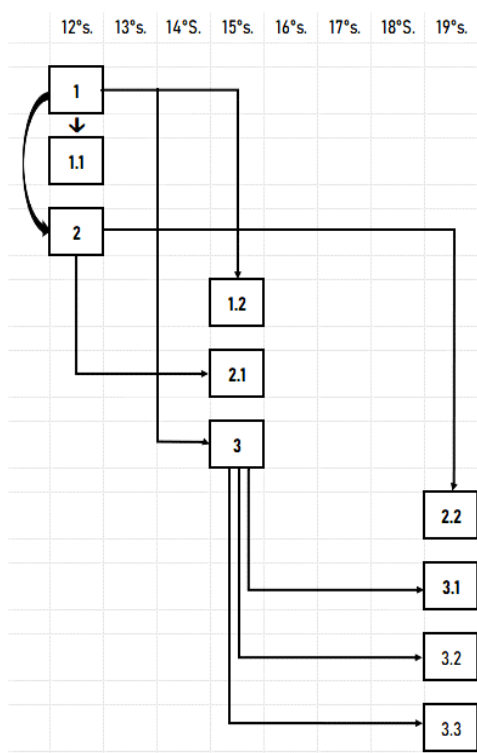


Figure 6 : Visualisation de la même polysémie évolutive par un graphe historique situé dans un cadre orthogonal {siècles \* sens}

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 (2020a) pour représenter la polysémie évolutive de l'adj. et n.m. *curieux*, cf. Fig. 7.



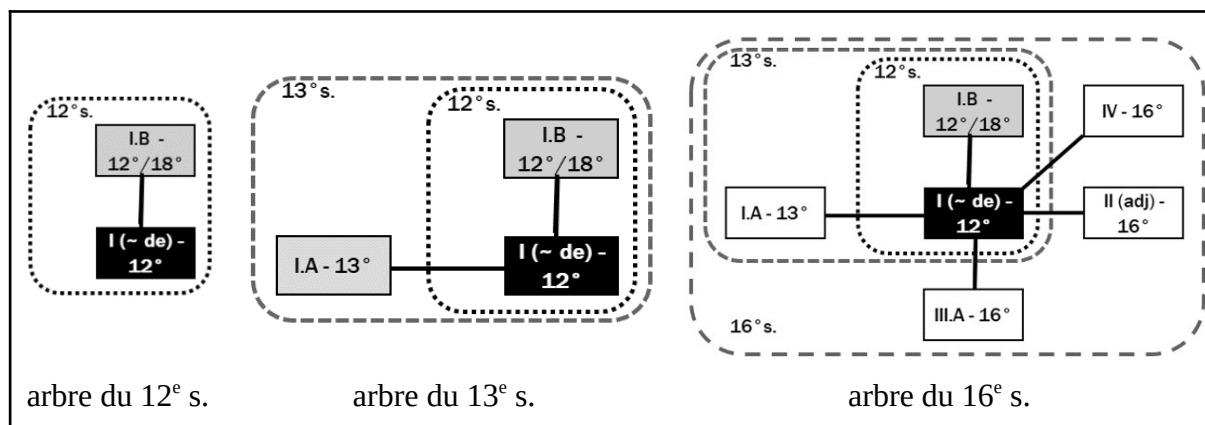


Figure 7 : Fragment de la visualisation de la polysémie évolutive de l'adj. et n.m. *curieux* par « instantanés séculaires » (Cf. François 2020a : 37-38)<sup>10</sup>

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 de stabilisation de la polysémie ne figurent pas. De ce fait – et aussi pour économiser de la place – nous n'avons pas encore cherché à automatiser cette technique. En revanche, la visualisation de la polysémie évolutive d'un vocable par un graphe historique du type suggéré dans la figure 6 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 d'un tableau de commentaires.
2. Et en outre on peut aisément apprécier l'ensemble des sens apparus respectivement au 12<sup>e</sup>, au 15<sup>e</sup> et au 19<sup>e</sup> 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 12<sup>e</sup>, le 16<sup>e</sup> et le 19<sup>e</sup> et non par exemple le 17<sup>e</sup>, 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

10 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.

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 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 14<sup>e</sup> siècle avec le *Dictionnaire françois-latin* de Robert Estienne (1539) suivi en 1610 du *Thresor de la Langue Francoyse* 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. 1.

### ■ 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 & 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).

### ■ 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 17<sup>e</sup> siècle ont une section historique. À titre d'exemple, celle – particulièrement copieuse – du n.m. *coeur* énumère 137 citations qui se répartissent ainsi :

11 <sup>e</sup>	12 <sup>e</sup>	13 <sup>e</sup>	14 <sup>e</sup>	15 <sup>e</sup>	16 <sup>e</sup>
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 16<sup>e</sup> 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 9<sup>e</sup> au 20<sup>e</sup> 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*<sup>11</sup>, dont les entrées bénéficient des recherches des éditions Le Robert complétées par de nombreux emprunts au TLFi.

Le diagramme de synthèse historique en Fig. 8 réunit deux stades historiques,

- a) d’abord les principaux dictionnaires de la langue française depuis leur apparition au début du 17<sup>e</sup> siècle et les dictionnaire spécialisés, étymologiques ou historiques, jusqu’à la finition du TLF,
- b) ensuite le TLF avec ses deux éditions successives, en 16 volumes et sur support électronique, avec ses deux principaux types d’entrée<sup>12</sup>.

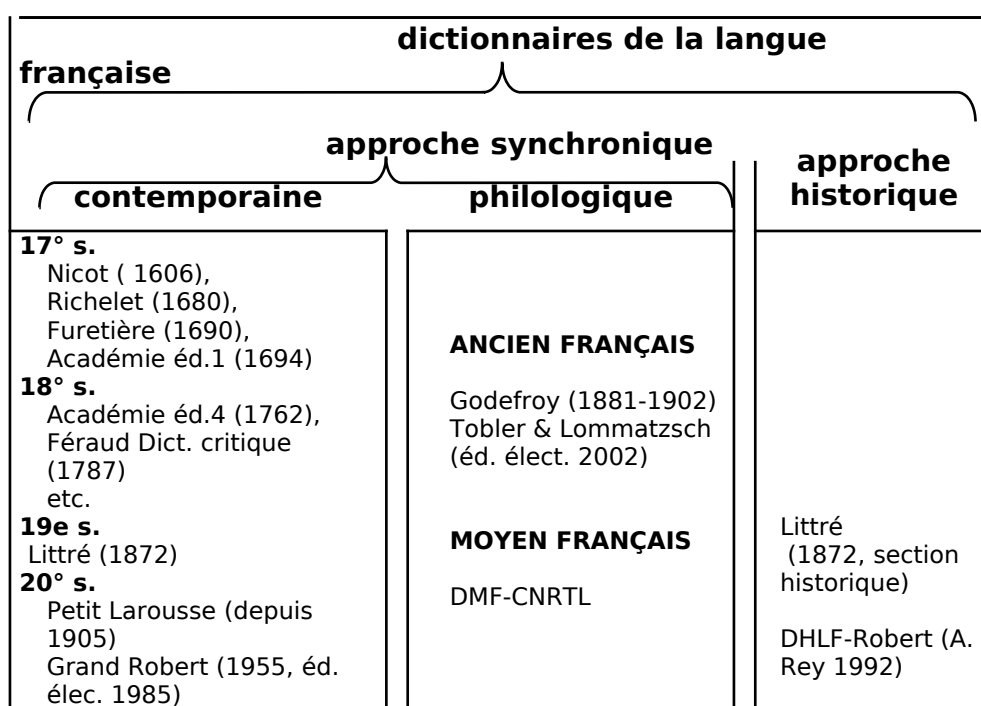


Figure 8 : Classement des dictionnaires antérieur au TLF et dont il a tiré parti, composition de ses articles et positionnement de notre double projet

Et le diagramme en fig.9 indique le positionnement de notre double projet dans le sillage du TLF dans son édition originale et du TLFi.

11 Le DHLF est paru en1992, sous la direction d’Alain Rey, et vient d’être réédité en 2022 pour son 30<sup>ème</sup> anniversaire.

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

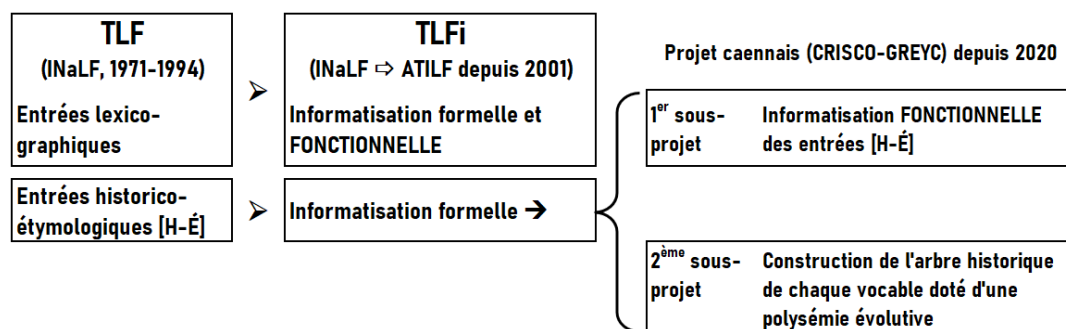


Figure 9 : Positionnement de notre projet dans le prolongement de l'informatisation du TLF

### 3. La microstructure des articles du TLFi et notamment de leur entrée historico-étymologique

#### 3.1. Articles, entrées, dégroupements et regroupements

Dans une présentation des caractéristiques techniques du TLFi et de leur exploitation, paru en 2010, Pascale Bernard, de l'ATILF, assimile<sup>13</sup> *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 dictionnaire est liée à une double pratique lexicographique :

- a) 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.

■ Entrez une forme

Chercher

[options d'affichage](#)

catégorie : toutes ▼

■ VASE<sup>1</sup>, subst. masc.

VASE<sup>2</sup>, subst. fém.

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

- b) D'un autre côté, l'inventaire des entrées lexicographiques est inférieur à la liste des lemmes en raison de **REGROUPEMENTS** morphosémantiques. Ainsi dans l'entrée « présent<sup>1</sup>, -ente » sont regroupés sous l'intitulé REMARQUES :

13 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 ».

**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)<sup>14</sup> sont implicitement subdivisées en une sous-entrée historique immédiatement suivie d'une sous-entrée étymologique. Ces entrées H-É se laissent classer en deux types :

- a) 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 5<sup>ème</sup> rang  $\alpha$  peut intervenir) et qui sont donc dotées d'une POLYSÉMIE ÉVOLUTIVE,
- b) 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.

<sup>14</sup> 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 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 entrées lexicographique et historico-étymologique, par la suite, constituent deux obstacles difficiles à contourner pour l'informatisation fonctionnelle de ces entrées.

### 3.2. Comparaison entre les deux entrées lexicographique et 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 (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.

<b>I</b>	[Le cœur dans sa réalité physique]
<b>A</b>	[Le cœur comme organe interne]
<b>B</b>	<i>P. méton.</i> Poitrine, qui abrite le cœur (et, secondairement, les autres organes internes primordiaux); <i>en partic.</i> l'endroit de la poitrine où les battements du cœur sont perceptibles
<b>II</b>	[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? » ( <i>Théol. cath.</i> t. 3, 11911).
<b>A</b>	[P. réf. à l'automatisme cardiaque; le cœur comme organe ou lieu d'une saisie plus ou moins automatique]
<b>B</b>	[P. réf. à l'intériorité et à l'activité de l'organe en tant que facteur central de la vie hum. individuelle] <i>Le cœur comme foyer ou réceptacle de la vie intérieure profonde, de la personnalité morale d'un individu.</i>
<b>C</b>	Le cœur comme foyer ou réceptacle du dynamisme moral, de certaines tendances volitives]
<b>D</b>	[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<sup>15</sup>

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 quatre pour II. Ces douze rubriques se décomposent à leur tour en 35 sections, soit une moyenne de  $\approx$  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 2 ci-dessous reproduisant la microstructure de l'entrée historico-étymologique du n.m. *cœur*, on voit que les rédacteurs de ces entrées ont fait un usage systématique de quatre types de caractères (romains, italiques, gras et petites capitales), pour spécifier différents types de données (dont l'éventail est cependant supérieur à celui des types de caractères, ce qui oblige à tenir compte également de la disposition des segments concernés par rapport aux identifiants de la microstructure des entrées et à des signes typographiques, notamment les parenthèses<sup>16</sup>)

La sous-entrée étymologique figure encadrée sur fond gris.

#### 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).

<sup>15</sup> Le surlignage a été effectué par l'ATILF à partir de l'informatisation fonctionnelle de l'entrée lexicographique.

<sup>16</sup> Voir plus bas pour l'interprétation contextuelle des types de caractère. Dans le site du CNRTL, la structuration de la version source en langage HTML a permis de surligner automatiquement différents types de données, chacun avec une couleur différente. Nous n'avons pas pu reproduire ici ces surlignages, mais on peut consulter la version superficiellement différente de cet article sous la forme du Cahier du CRISCO n° 37 *La polysémie évolutive du lexique français (XIIe-XXe siècle) : Projet d'informatisation fonctionnelle et de modélisation graphique des entrées historiques du TLFi*, consultable par le portail HAL.

[I.]B. Ca 1100 au propre (*Roland*, éd. J. Bédier, 2965 : [Li emperere ad fait] tuz les **quers** en paille recueillir).  
 [I.]C. Ca 1100 p. ext. « la poitrine » (*ibid.*, 3448 : L'escut li freint, cunter 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);  
 XIII<sup>e</sup>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 XII<sup>e</sup>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**);  
 1<sup>er</sup> tiers XIII<sup>e</sup>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 XIV<sup>e</sup>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 XVII<sup>e</sup>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. *cōr* (peut-être par l'intermédiaire d'une forme \**cōre*, FOUCHÉ, p. 656, BL.-W.<sup>1-5</sup>) 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. κ α ρ δ ι ' α « 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)

L'interprétation fonctionnelle des types de caractère et des signes typographiques procède comme suit :

- les chaînes 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** ;
- les chaînes de caractères entre guillemets véhiculent une **DÉFINITION** ;
- les chaînes de caractères introduites par < : > dans un espace entre parenthèses et comportant la vedette en caractères gras véhiculent des **CITATIONS** ;
- les chaînes de segments en caractères romains figurant immédiatement après un identifiant de classe superordonnée, véhiculent une 'PRÉ-DÉFINITION' (ex. **II. Centre de la vie intérieure**).

En même temps, cet exercice visant à préfigurer l'informatisation fonctionnelle des entrées H-É, fait ressortir des obstacles qu'il importe de contourner à cette fin, 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<sup>e</sup> (celui des premiers manuscrits enregistrés dans **FRANTEXT**) et le 20<sup>e</sup> (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 (police par défaut) et dont seules 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.

#### 4. L'éventail typographique des entrées historico-étymologiques [H-É]

L'informatisation fonctionnelle des entrées H-É peut s'effectuer pas à pas à l'aide de différents outils dont seule la 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 (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 figurent 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 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<sup>17</sup>, ex. (entrée PARTICULARISME, aspect de l'édition publique et notation XML<sup>18</sup>) :

<b>Étymol. et Hist.1.</b> 1689 théol.	<code>&lt;etymology&gt;&lt;hi rend="bold"&gt;Étymol. et Hist.1.&lt;/hi&gt; 1689 théol. (...)</code>
--	---

- [2] La seule chaîne de caractères immédiatement à droite figure en gras. Dans ce cas, le point sépare la fin d'une rubrique de 1<sup>er</sup> 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.

p.176). <b>II.</b> Subst. <b>A.</b> masc	(...) <code>&lt;hi rend="bold"&gt;II.&lt;/hi&gt;</code> (...)
--	---

- [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 eux-ci, ex. (article PARTITION), ex.

<b>PARTITION</b> , subst. fém. <b>Étymol. et Hist.I. 1.</b> Ca 1175	<code>&lt;hi rend="bold"&gt;Étymol. et Hist.I. 1.&lt;/hi&gt;</code>
---	---

- [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 <i>parvenir</i> *.	Part. passé de <code>&lt;hi rend="italic"&gt;parvenir*&lt;/hi&gt;</code>
-----------------------------------	--

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) :

<sup>17</sup> 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

<sup>18</sup> 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.

**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<sup>e</sup>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.)]

#### ■ 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 variables selon leur encadrement typographique :

FONCTION CONTEXTUELLE DES TYPES DE CARACTÈRES	EN DEHORS DE PARENTHÈSES COURBES	ENTRE PARENTHÈSES COURBES
Caractères gras Caractères italiques Petites capitales	Identifiants Forme de mot ou collocation Numérotation des siècles en chiffres romains	Vedette dans une citation Titre d’une référence Auteur d’une référence

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

Ainsi par exemple la mise en forme de la rubrique 2. met en œuvre dix segments en caractères romains, gras et italiques avec deux espaces entre parenthèses courbes.

**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)

<hi rend="bold">2.</hi> indique un changement d'état, d'aspect <hi rend="bold">a) </hi><hi rend="italic">ca</hi> 1165 <hi rend="italic">torner a porreture</hi> (en parlant d'un cadavre) (<hi rend="smallcaps">Benoit de</hi> <hi rend="smallcaps">Ste</hi>-<hi rend="smallcaps">Maure</hi>, <hi rend="italic">op. cit.</hi>, 22397)

Le tableau 4 explicite la fonction de chacun de ces segments :

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

Tableau 4 : Format XML des segments figurant en caractères gras, italiques et petites capitales et fonction des types de segments ainsi identifiés

#### ■ 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. Bilan d'étape et illustration d'un champ d'étude dérivé

### 5.1. 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 (à droite dans la fig. 10) 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 gauche 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<sup>e</sup> siècle d'où dérive un nœud générique reconstitué [2.] au 17<sup>e</sup> s., lequel a trois nœuds-fils [2a], [2b] et [2c] respectivement au 16<sup>e</sup>, 18<sup>e</sup> et 20<sup>e</sup> s.

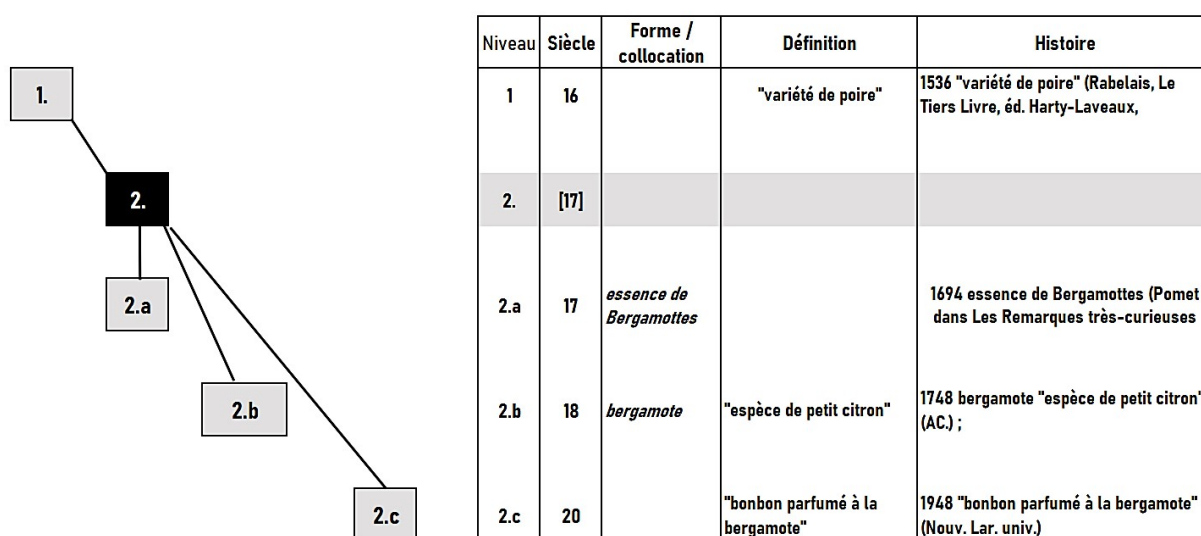


Figure 10 : Modélisation graphique de la polysémie évolutive du n.f. BERGAMOTE

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] à [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 nœud-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'enchaînement 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] (enchaînement) 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 à 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 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/c].

## 5.2. *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<sup>19</sup>, 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 1<sup>ère</sup> attestation de chaque nouveau type d'emploi), on peut s'attendre à ce que la fréquence relative<sup>20</sup> 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 attestés 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*
- 3 verbes : *casser, jeter, monter*

et pour chaque vocable nous avons comparé le nombre cumulé des rubriques de leur entrée historique siècle par siècle et la fréquence relative du vocable dans chaque corpus séculaire de sa première attestation jusqu'au 20<sup>e</sup> siècle. Pour que ces deux courbes figurent aisément sur un même graphique, la fréquence relative a été multipliée par 10<sup>5</sup>. 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.

19 Voir l'analyse historique de la polysémie du n.m. *timbre* par Darmesteter (1887 : 74-5) qui n'est plus plausible au 21<sup>e</sup> 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).

20 C'est-à-dire le nombre des occurrences du vocable de siècle en siècle divisé par le volume de chaque corpus séculaire (également désigné comme la « présence séculaire » [PrsSéc] du vocable dans FRANTEXT).

**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* et *jeu* et pour le verbe *casser*, l'orientation des deux tendances est apparentée : ↗↗ et aucune fluctuation n'a perturbé la tendance croissante de la présence séculaire des trois vocables.

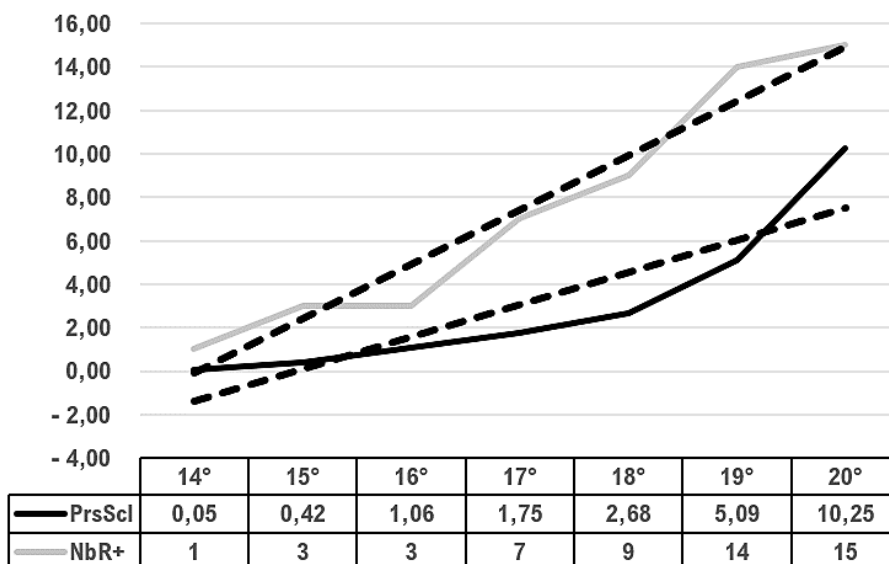


Figure 11 : Profil historique couplé du n.f. *carte*

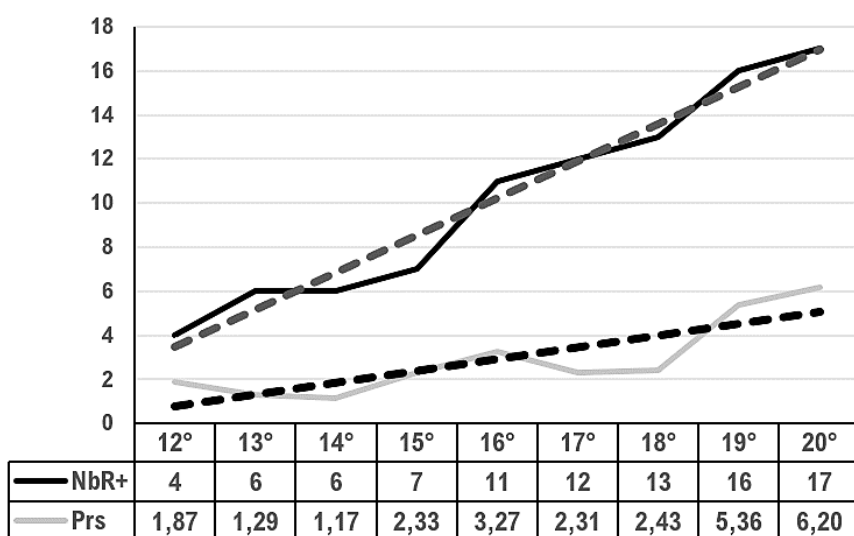


Figure 12 : Profil historique couplé du v. *casser*

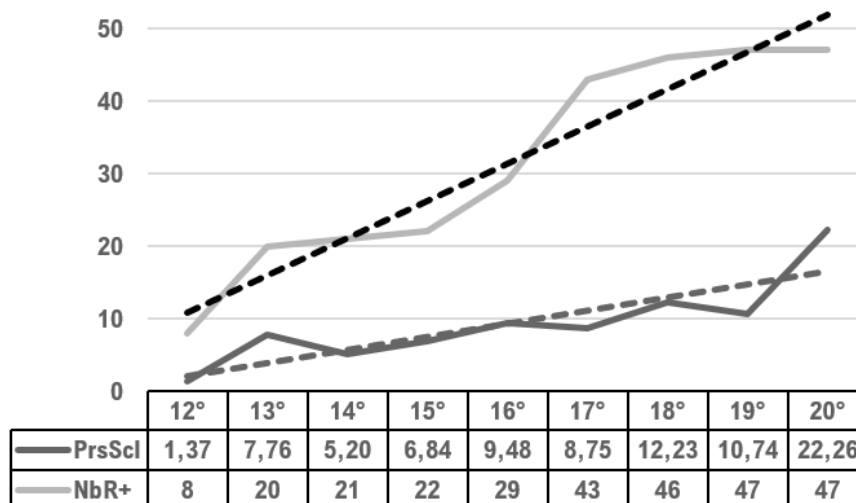


Figure 13 : Profil historique couplé du n.m. *jeu*

**Classe II :** Concordance globale entre la droite de tendance du nombre cumulé des rubriques et celle de la présence séculaire du vocable dans FRANTEXT au prix de fluctuations  $\square$  *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 deux courbes sont apparentées :  $\nearrow \nearrow$ , mais on constate des fluctuations de la présence séculaire du vocable.

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

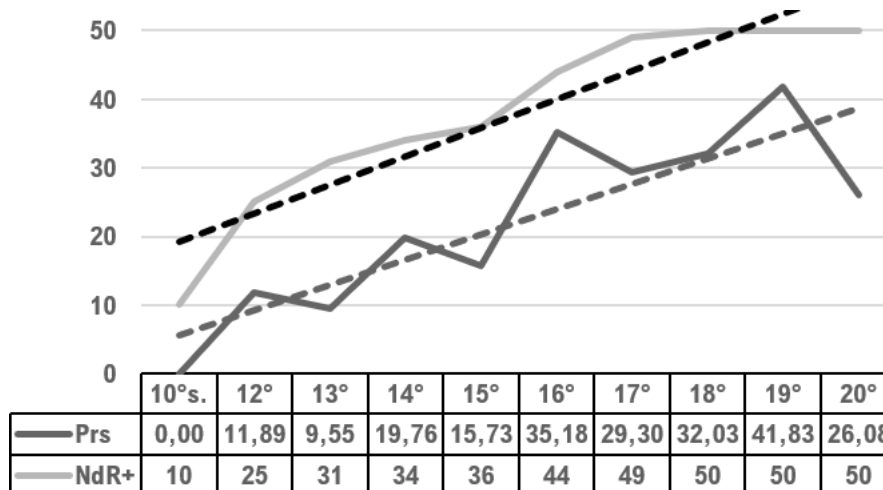


Figure 14: Profil historique couplé du v. *jeter*

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

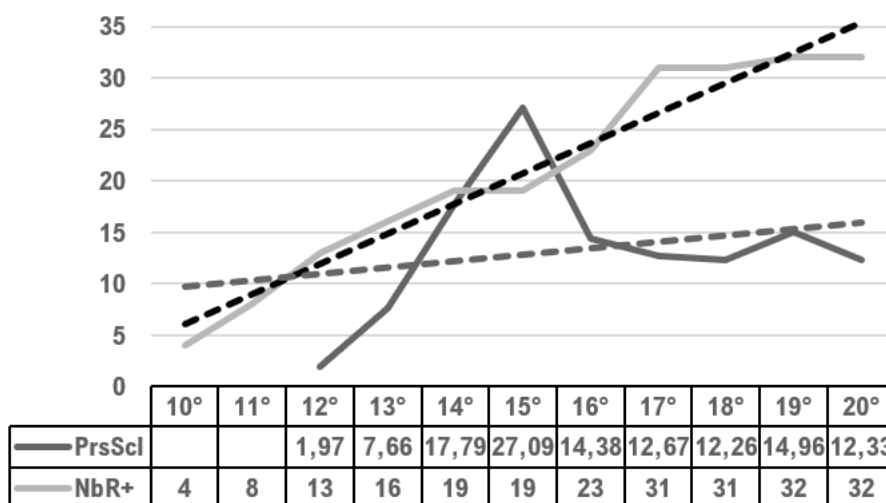


Figure 15 : Profil historique couplé du n.m. *jour*

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

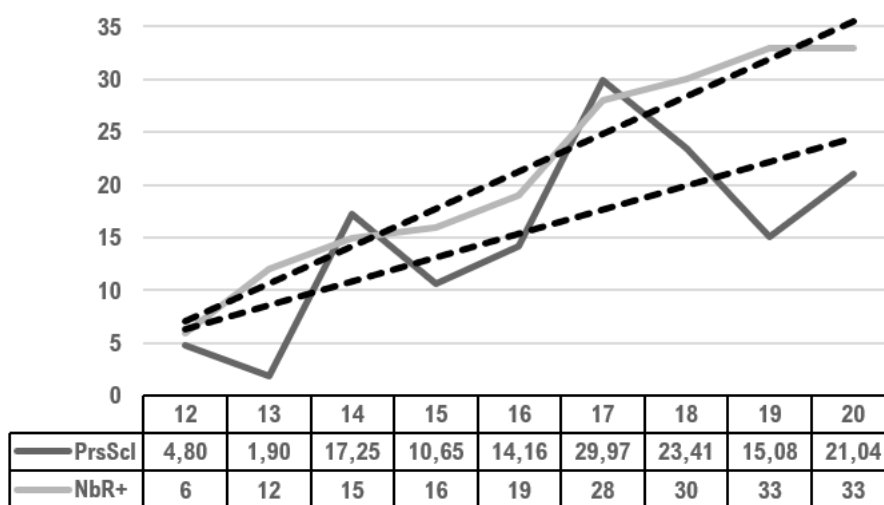


Figure 16 : Profil historique couplé de l'adj. et n.m. *juste*

La courbe de présence séculaire de l'adj. et n.m. *propre* 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 15<sup>e</sup> et le 16<sup>e</sup> siècle) et avant une aussi brusque décroissance entre le 18<sup>e</sup> et le 19<sup>e</sup> (de 41,76 à 23,15).



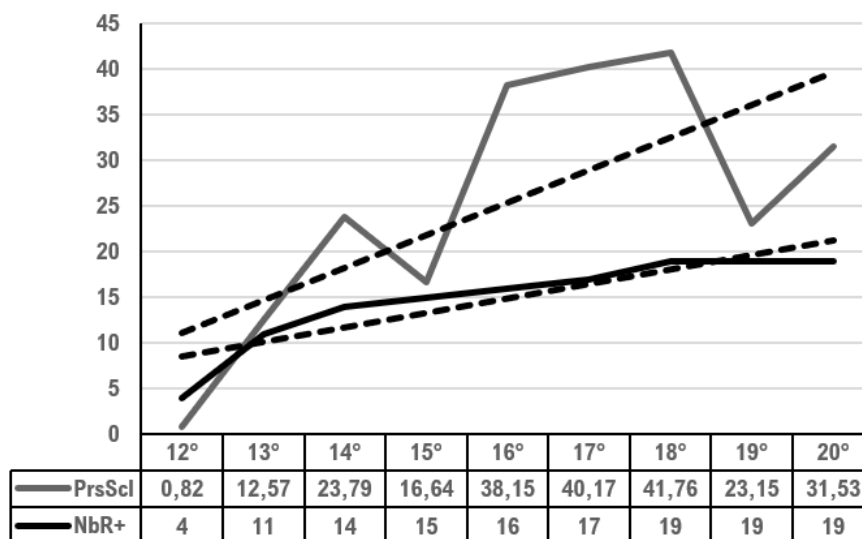


Figure 17 : Profil historique couplé de l'adj. et n.m. *propre*

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

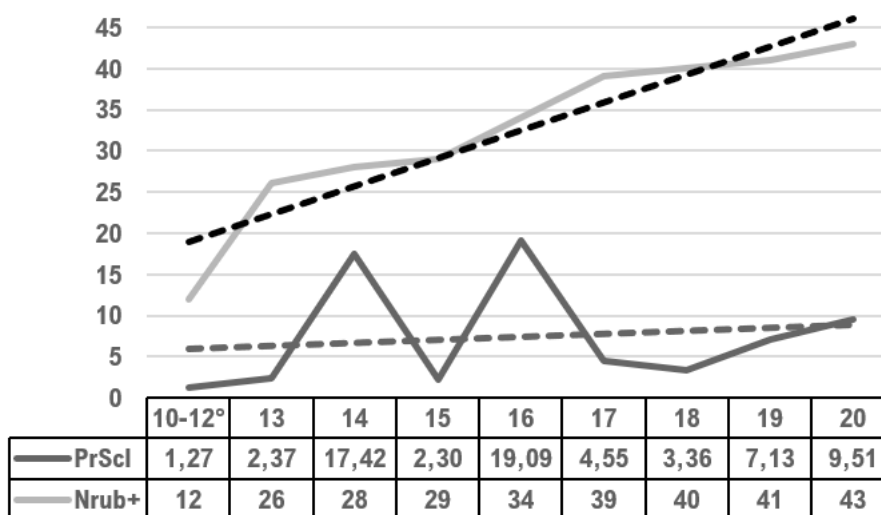


Figure 18 : Profil historique couplé de l'adj. *sec*

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 vocables 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 très élevées pour le 12<sup>e</sup> s. (32,66) et surtout le 13<sup>e</sup> s. (46,44) peuvent sans doute s'expliquer par l'ambivalence de la forme **clers** qui désigne fréquemment un *clerc* (<lat. *clericus*) non l'adj. *clair* (<*clarus*). Cette ambivalence vaut jusqu'au 13<sup>e</sup> s., ce qui pourrait expliquer qu'à partir de sa disparition au 14<sup>e</sup> 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 13<sup>e</sup> 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.

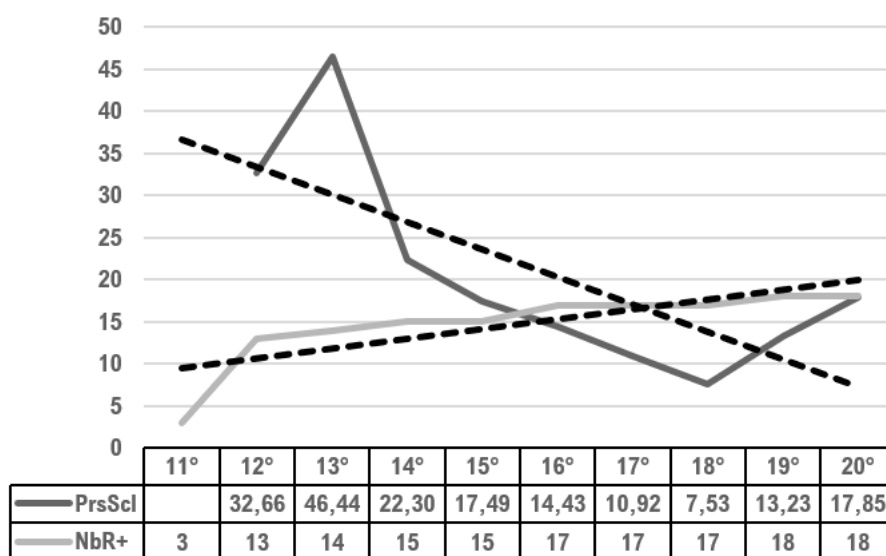


Figure 19 : Profil historique couplé de l'adj. et n.m. *clair*

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

Mais la concurrence entre *visible* et *clair* diminue à partir du 18<sup>e</sup> siècle et finit par se stabiliser entre 25 et 30%. Au-delà du 17<sup>e</sup> siècle, aucun des deux facteurs avancés ne peut donc éclairer la décroissance et c'est l'époque où la concurrence entre *visible* et *clair* diminue presque de moitié (cf. fig. 20).

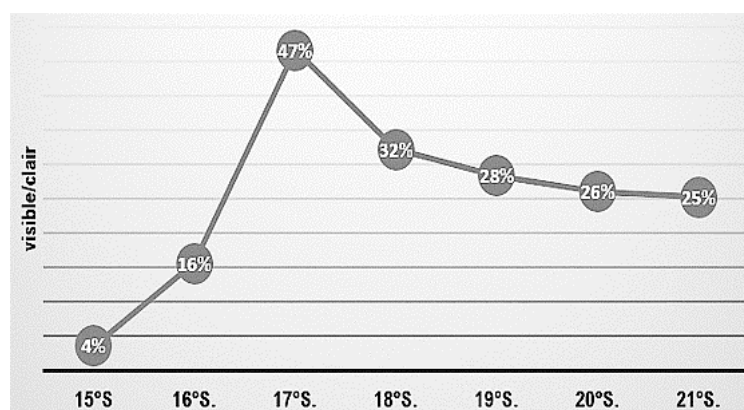


Figure 20 : Rapport entre les deux fréquences  
de **visible** et de **clair** entre le 15<sup>e</sup> et le 21<sup>e</sup> siècles

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

- a) entre le 13<sup>e</sup> et le 14<sup>e</sup> siècle, la disparition de l'homographie entre *clers\_1*(le clerc) et *clers\_2* (l'adj. *clair*)
- b) et entre le 15<sup>e</sup> et le 18<sup>e</sup> s. la concurrence entre les adj. *clair* et *visible* (*transparent* et *perceptibles* 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 15<sup>e</sup> et le 18<sup>e</sup> 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. 21 et François 2010). Un seul changement sociétal remarquable vint 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.

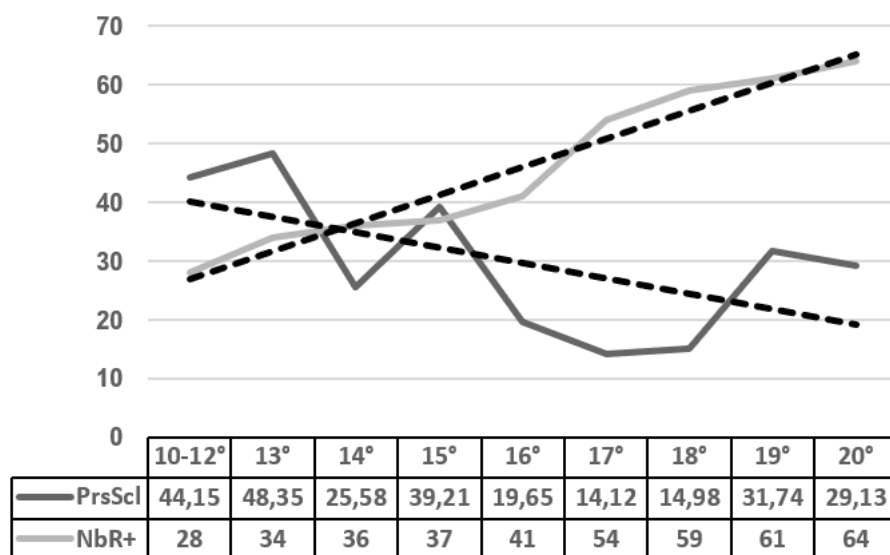


Figure 21 : Profil historique couplé du v. *monter*

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.22) :

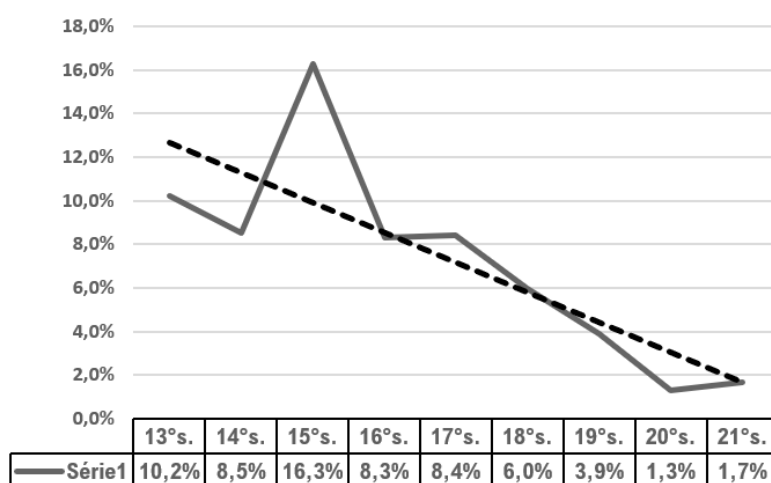


Figure 22 : Rapport entre les deux fréquences de *monter*  
et *monter* [...max.3 mots...] entre le 13<sup>e</sup> et le 21<sup>e</sup> siècles

Si l'on compare l'orientation de la droite de tendance dans les deux figures 21 et 22, 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. *monter*, 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 15<sup>e</sup> 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 :

- a) 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 :
- b) 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.

## 6. La poursuite de l'informatisation des entrées H-É

Dans les sections précédentes, nous avons envisagé la modélisation graphique des entrées H-É du TLFi comme l'objectif privilégié et leur informatisation fonctionnelle comme le moyen d'y parvenir. Cependant l'élaboration du graphe historique arborescent de chaque entrée sur la base du tableau des données qui l'accompagne doit tenir compte également de contraintes cognitives chez leurs lecteurs. L'arbre historique du n.f. BERGAMOTE, que nous avons présenté en fig. 10, est élémentaire, avec ses cinq nœuds et ne donne pas une image réaliste de la complexité que peuvent présenter les arbres historiques de vocables qui peuvent comporter un nombre bien supérieur de nœuds. Il devient rapidement impossible de réunir dans un même écran l'intégralité des données supportant la composition de l'arbre, même en limitant (comme nous le faisons déjà) la visualisation primaire des rubriques à une seule ligne que le lecteur peut dérouler selon son intérêt privilégié. Il reste donc à mettre en place une technique de présentation réduite de l'ensemble de l'arbre combinée à la présentation développée de chaque groupe de rubriques.

En outre, comme nous l'avons souligné au long de cet article, l'informatisation fonctionnelle des rubriques de chaque entrée historique est confrontée à un défi au-delà de l'interprétation fonctionnelle des segments fondée sur la typologie des caractères employés et des signes typographiques. À titre d'exemple, l'identification des définitions classiques entre guillemets ne présente aucune difficulté, mais celle des « quasi-définitions » dénuées de guillemets est plus délicate, car elle implique le repérage de formulations répétitives caractéristiques, comme « indique + GN / que P ». Par ailleurs, les rédacteurs de ces entrées H-É ont reçu dès le début des années 1970 la consigne de ne pas répéter une définition quand elle s'applique à deux rubriques de suite. Dans ce cas, la justification de l'ouverture d'une nouvelle rubrique peut-être la variation de la forme du mot ou la prise en compte d'une collocation ou encore la spécifica-

tion d'une condition d'emploi qui est véhiculée par une métadonnée de domaine conceptuel (par ex. celui de la marine), de registre (par ex. argotique), de syntaxe (par ex. de préposition vs postposition) ou de sémantique (par ex. une dérivation par analogie, métaphore ou métonymie). Ces spécifications figurent en caractères romains, mais, comme l'objectif éditorial principal était d'occuper le moins de place possible, il s'agit essentiellement d'abréviations et l'établissement du catalogue de ces spécifications bénéficie de la liste des abréviations fournie dans l'introduction au premier volume de l'édition papier du TLF.

Un tel affinement de l'identification des données fonctionnelles n'est pas indispensable à la modélisation graphique des entrées historiques. Il est à envisager dans une optique inverse, en renversant les priorités au profit de l'exploitation transversale de ces entrées. On peut dès maintenant comparer le contenu des définitions successives de chaque rubrique d'une entrée, mais en l'absence de définition classique, il faut pouvoir repérer des données substitutives fournies soit par des quasi-définitions, soit par des métadonnées indiquant la spécification de la définition fournie par la rubrique précédente<sup>21</sup>, lesquelles se présentent toutes en caractères romains et avec pour tout indice typographique le point d'abréviation. Ce projet sera donc poursuivi en le réorientant en direction de l'affinement des types de données afin de permettre aux chercheurs en lexicologie historique de disposer d'outils d'analyse transversale comparables à ceux qui font des entrées lexicographiques du TLFi une mine d'informations lexicales, grammaticales et littéraires effectivement exploitées depuis un quart de siècle.

## ANNEXE STATISTIQUE ET INFORMATIQUE

### 1. Données statistiques

#### 1.1. Entrées historiques monosémiques vs polysémiques par siècle

L'enregistrement des 49 862 entrées H-É du TLFi a révélé que les entrées historiques monosémiques<sup>22</sup> représentent 58% du total (29 008 contre 20 854 pour les polysémiques, comme indiqué sur l'interface web<sup>23</sup>). Le tableau 1 répertorie le nombre d'entrées monosémiques par siècle (cf. figure 1) et, concernant les entrées polysémiques, pour chaque siècle le nombre d'entrées dont la première ou la dernière rubrique est datée durant ce siècle.

Siècle	9 <sup>e</sup>	10 <sup>e</sup>	11 <sup>e</sup>	12 <sup>e</sup>	13 <sup>e</sup>	14 <sup>e</sup>	15 <sup>e</sup>	16 <sup>e</sup>	17 <sup>e</sup>	18 <sup>e</sup>	19 <sup>e</sup>	20 <sup>e</sup>	NC
Nb d'entrées M	1 1	47	14 3	15 91	15 50	177 2	13 92	35 27	24 69	34 95	79 49	35 22	15 40
Nb d'entrées P - première rubrique	6 1	28 6	39 7	36 90	25 27	244 8	15 81	29 26	15 06	14 76	21 65	43 3	13 50
Nb d'entrées P - dernière rubrique	3	10	13	56 7	67 3	661	59 8	21 30	31 71	28 94	66 11	33 07	20 8

21 Il arrive bien entendu qu'une rubrique comporte à la fois une définition classique et une métadonnée de domaine, registre, syntaxe et/ou sémantique, auquel cas les deux informations se combinent.

22 Nous définissons comme « entrée historique monosémique » toute entrée qui ne donne pas lieu à l'introduction de rubriques. Toutes les autres présentent plusieurs rubriques et sont à ce titre des « entrées historiques polysémiques ». Il est à noter que l'identification alphanumérique des rubriques de l'entrée principale (ou « lexicographique ») et celle des rubriques de l'entrée historiques sont indépendantes l'une de l'autre.

23 Cf. < <https://crisco3.unicaen.fr/noticestlfi/> >

Tableau 1 : Entrées monosémiques [M] et polysémiques ([P], datation de la première et de la dernière rubrique) entre le 9<sup>e</sup> et le 20<sup>e</sup> siècles<sup>24</sup>.

Le graphique en fig.1 montre qu'une grande majorité des entrées monosémiques sont attestées au 19<sup>e</sup> siècle puis ensuite quasiment à égalité aux 16<sup>e</sup>, 18<sup>e</sup> et 20<sup>e</sup> siècle. La barre du 19<sup>e</sup> (7949 entrées) est plus longue que les deux barres cumulées du 18<sup>e</sup> (3495) et du 20<sup>e</sup> siècle (3522). Le doublement du nombre des entrées monosémiques entre le 18<sup>e</sup> et le 19<sup>e</sup> s. est assurément corrélé à l'enregistrement, au 19<sup>e</sup> siècle, d'un grand nombre de nouveaux vocables dans les grands dictionnaires de P. Larousse (1869), L.N. Bescherelle (1856), É. Littré (1872) et A. Hatzfeld & A. Darmesteter (1890), ce qui leur a permis de se diffuser dans la littérature scientifique et technique, les journaux et les belles-lettres. Entre le 19<sup>e</sup> et le 20<sup>e</sup> siècle, ces vocables n'ont pas eu le temps de développer une polysémie évolutive et ils figurent donc comme monosémiques dans leur entrée H-É.

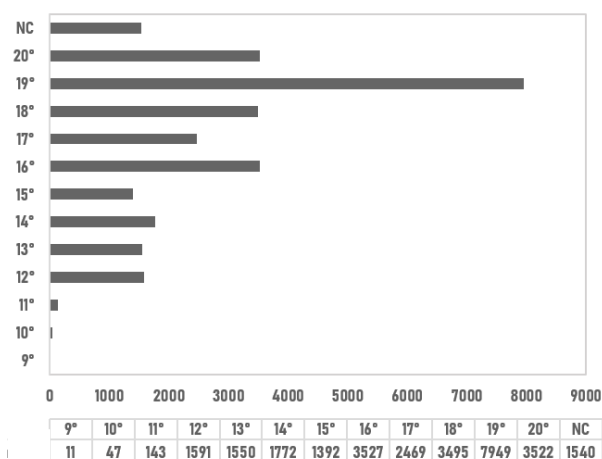
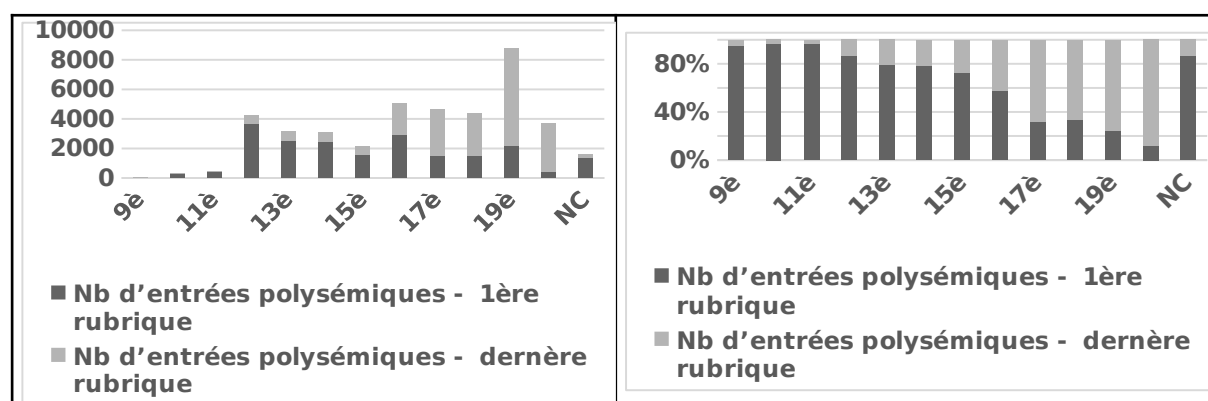


Figure 1 : Graphique en barres du nombre des entrées historiques monosémiques par siècle

Les deux graphiques en figure 2 se prêtent à une double lecture : à gauche on trouve la fréquence ABSOLUE des entrées polysémiques dont la première rubrique (en bas) ou la dernière (en haut) figure dans chaque siècle ; à droite, il s'agit de leur fréquence RELATIVE.



Histogrammes empilés

Histogrammes empilés à 100%

Figure 2 : Entrées H-É polysémiques classées par siècle de 1<sup>ère</sup> attestation

Entre le 9<sup>e</sup> et le 11<sup>e</sup> siècles, les entrées dont la première rubrique est datée entre le 9<sup>e</sup> et le 11<sup>e</sup> siècles sont très peu nombreuses et aucune dernière rubrique n'est datée de cette époque. Du 12<sup>e</sup> au 15<sup>e</sup> siècle, la proportion de premières rubriques est très supérieure à celle des dernières,

24 NC : « non connu »

comme on pouvait s'y attendre, puisqu'à cette époque les polysémies évolutives commencent à se bâtir. Inversement, du 16<sup>e</sup> au 20<sup>e</sup> siècle la proportion des dernières rubriques s'accroît graduellement et c'est au 20<sup>e</sup> siècle qu'elle est la plus élevée, ce qui indique que le nombre des types d'emploi nouveau diminuent et donc que l'entrée perd progressivement son pouvoir d'attraction pour de nouvelles extensions de sens. L'accroissement relatif des dernières rubriques est plus explicite dans le graphique de droite. Mais ce que révèle celui de gauche, c'est la place remarquable du 16<sup>e</sup> et du 19<sup>e</sup> siècle. Au 17<sup>e</sup>, la fréquence absolue des dernières rubriques se rapproche de celle des premières, c'est-à-dire qu'une part notable du lexique polysémique (encore retenu par le TLFi, c'est-à-dire encore en usage à la fin du 20<sup>e</sup> siècle, au moins dans certaines langues de spécialité) n'a plus donné lieu à de nouvelles extensions de sens au-delà au 16<sup>e</sup> siècle, une fois que la révolution de l'imprimerie s'est imposée et à prodigieusement accru la diffusion des écrits. Au 19<sup>e</sup>, le nombre des premières rubriques (et donc des nouveaux vocables) est supérieur à celui des deux siècles précédents, mais celui des dernières rubriques (et donc de la clôture de polysémies évolutives) équivaut aux clôtures polysémiques cumulées des 17<sup>e</sup> et 18<sup>e</sup> siècles. L'interprétation de ces chiffres n'est pas évidente, mais on peut y voir un effet conjugué

- de l'enregistrement du lexique courant et technique depuis la fin du 17<sup>e</sup> siècle, avec deux jalons essentiels, la parution du *Dictionnaire universel* de Furetière (1790) en amont et celles du *Dictionnaire de la langue française* de Littré en 1872 et du *Dictionnaire général* de Hatzfeld et Darmesteter en fascicules à partir de 1890 en aval,
- et de l'accroissement de la scolarisation et de la pratique du « français national »<sup>25</sup> depuis la Révolution française (d'ailleurs les rédacteurs du TLF considèrent dans l'introduction du premier volume que la Révolution française a ouvert une « synchronie étendue » qui vaut donc depuis plus de deux siècles).

## 1.2. Statistiques sur les rubriques et leur niveau hiérarchique

### ■ NOMBRE D'ENTRÉES EN FONCTION DU NOMBRE DE RUBRIQUES

Le graphique en figure 3 représente le nombre d'entrées historiques (mono- et polysémiques) par groupes de cinq rubriques (1-5, 6-11, etc. jusqu'à 25-30). Compte tenu de la décroissance rapide du nombre d'entrées (en abscisses) pour chacun de ces six groupes (en ordonnées), l'échelle des abscisses est logarithmique. Mais l'effectif réel de chaque groupe est mentionné sur la barre correspondante.

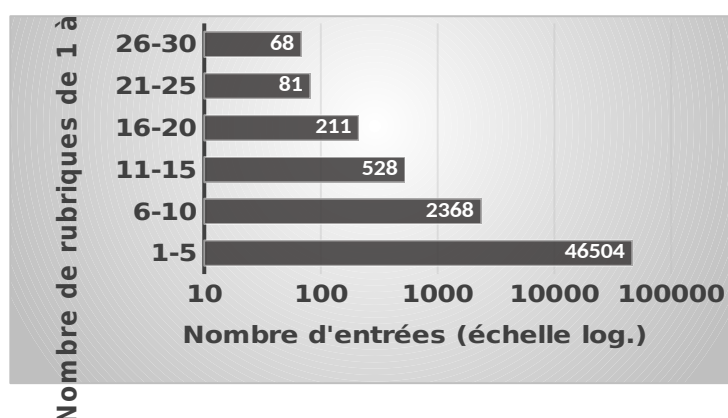


Figure 3 : Nombre d'entrées historiques présentant de 1 à 5 rubriques, 6 à 10, etc. jusqu'à 26-30

25 M. Bescherelle « aîné » a publié en 1856 la première livraison de son *Dictionnaire NATIONAL*, également qualifié en sous-titre d'*UNIVERSEL*, comme celui de Furetière 166 ans auparavant, ce qu'on peut interpréter comme « le dictionnaire qui traite de tout ce qui est exprimable et définissable (la dimension universelle) dans la langue de la nation française (la dimension nationale) ».



Nous constatons que le premier groupe comptant, de 1 à 5 rubriques, rassemble environ 12 fois plus d'entrées que le second qui en compte de 6 à 10 (46 504 vs 2 368). À son tour ce second groupe compte environ quatre fois plus d'entrées que le troisième comptant de 11 à 15 rubriques (538). Progressivement la décroissance se tasse, le dernier groupe (de 26 à 30 rubriques) accusant une régression de seulement 16% par rapport à l'avant-dernier (de 21 à 25 rubriques). Ce que l'échelle logarithmique ne met pas clairement en valeur, c'est qu'à lui seul le premier groupe (de 1 à 5 rubriques) représente 93% des 49 862 entrées au total. Ces 46 504 entrées se subdivisent à leur tour en 29 008 entrées monosémiques (62%) et 17 496 entrées faiblement polysémiques (de 2 à 5 rubriques, 38%). Au-delà on rencontre quelques entrées dotées d'un nombre exceptionnel de rubriques :

60	61	62	68	75	84	89	93	158
<i>porter</i> _1	<i>touche</i> r_1	<i>raison</i> rien	<i>tant</i>	<i>tourner</i>	<i>tête</i>	<i>que</i> e	<i>tomber</i> r_1	<i>tenir</i>

#### ■ LE DEGRÉ VARIABLE DE HIÉRARCHISATION DES ENTRÉES HISTORIQUES

Comme nous venons de le voir, le nombre des rubriques des entrées historiques varie de 1 pour 29 008 entrées, à 154 pour l'entrée du v. *tenir*<sup>26</sup>. La microstructure des entrées historiques doit tenir compte à la fois de classes d'emploi sémantiquement très distantes, de la succession chronologique d'extensions de sens dans des domaines conceptuels variés et de fluctuations syntaxiques corrélées à un éventail de micro-sens contextuels occasionnellement véhiculés par des collocations et des locutions.

Ces différents critères ont poussé les rédacteurs de ces entrées à les hiérarchiser au point de prévoir cinq niveaux d'affinement graduel. Nous avons repéré cinq niveaux de hiérarchisation de l'identifiant de chaque entrée. Le classement de niveau 1 figure en chiffres romains (I, V, X), celui du niveau 2 en majuscules latines (A, B, ..., Z), celui du niveau 3 en chiffres arabes (1, 2, ..., ∞), celui du niveau 4 en minuscules (a, b, ..., z) et celui du niveau 6 en minuscules grecques (α, ..., ω). On constate que 46% des 29 008 entrées polysémiques se contentent d'un seul niveau, 20% ont recours à deux niveaux et au final 2% tirent parti des cinq niveaux identifiés (notamment *part\_2*, *marquer*, *période*, *toucher*, *prétendre* ou *vider*).

#### ■ L'ÉVOLUTION DE LA PRIORITÉ DONNÉE AUX NIVEAUX DE STRUCTURATION SUPÉRIEURS VS INFÉRIEURS AU FIL DES 81 FICHIERS XML

Enfin, les données statistiques dont nous disposons désormais nous ont permis de constater une évolution assez mystérieuse de la méthode de hiérarchisation des entrées H-Ê appliquée par leurs rédacteurs au fil des décennies. Les entrées qui nous ont été confiées par l'ATILF couvrent 81 fichiers appliquant aux 49 862 entrées (dans l'ordre alphabétique des lemmes) un traitement informatique strictement formel en langage XML. Dans la première moitié de la série des 81 fichiers<sup>27</sup>, l'usage du premier niveau hiérarchique (I, II, III, etc.), élevé au début ne cesse de décliner, tandis que le cinquième niveau (α, β, γ, etc.) reste hors-jeu. En revanche, dans la seconde moitié de la série, ce niveau entre en jeu et dès le fichier 54, son emploi supplante celui du premier niveau.

26 Cf. François (1980) sur la polysémie du v. *tenir*.

27 Le fichier 40, situé au milieu de la série, va du n.f. *incartade* au n.m. *insigne*.

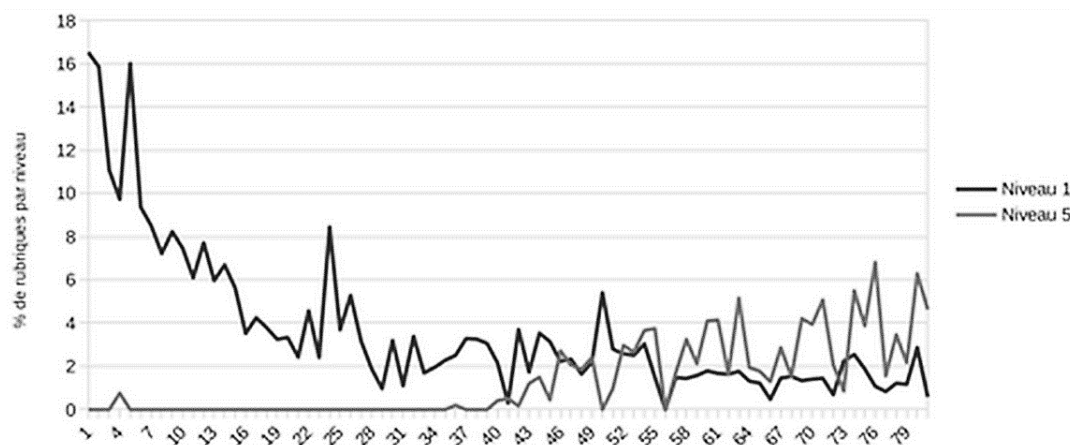


Figure 4 : La régression de l'usage du niveau I de classement des entrées historiques et la progression de celui du niveau  $\alpha$  au fil de la série des 81 fichiers en langage XML

Concrètement, cela signifie que – quel que soit l'empan hiérarchique des identifiants les plus fins, par exemple l'empan 3 pour un identifiant du type **A.1.a** – de nombreux identifiants de la première moitié de la série tirent parti du rang hiérarchique supérieur, sans qu'aucun ne recoure au rang inférieur et que, dans la seconde moitié de la série, la majorité des identifiants ne recourent pas au rang supérieur, mais tirent parti du rang inférieur.

A priori, il n'y a aucune raison pour que la distance entre les groupes d'emploi des vocables polysémiques varie entre les fichiers 1-40 (A-I) et les fichiers 41-81 (J-Z). Et l'évolution est trop régulière (surtout dans le premier et le dernier quarts) pour avoir un caractère aléatoire. Une seule explication subsiste : les rédacteurs ont probablement reçu – dans les années 1980 ? – la consigne de limiter le recours au degré d'écart sémantique représenté par le niveau supérieur et d'exploiter plus régulièrement le niveau inférieur. Mais dans la mesure où cette hiérarchie des écarts sémantiques ne peut se fonder que sur un jugement intuitif de distance sémantique décroissante de  $\overline{\text{I}}$  à  $\overline{\text{A}}$ , à  $\underline{\text{I}}$ , à  $\underline{\text{a}}$  et finalement à  $\alpha$ , l'intérêt de cette consigne présumée reste douteux.

## 2. Les deux formats des entrées historiques, publique et en codage XML

Dans la section 4, l'éventail typographique des entrées historico-étymologiques du TLFi a été présenté avec de nombreux extraits du TLFi. Ces extraits comportaient deux « vues ». La première vue est celle de l'internaute naviguant sur le site du TLFi, il s'agit d'un texte mis en forme. La seconde vue est celle du code XML, qui est le format dans lequel sont stockées toutes les entrées du TLFi. Lorsqu'un internaute se connecte sur le site du TLFi et recherche un mot, le serveur (c'est-à-dire l'ordinateur sur lequel est hébergé le TLFi) extrait du fichier XML la partie qui concerne le mot recherché et le transforme pour en faire une page web.

Les 81 fichiers XML sur lesquels nous travaillons comportent l'intégralité du texte des notices H-É du TLFi. Le découpage en 81 fichiers est probablement dû à la volonté de limiter la taille des fichiers au moment de leur création. Un fichier XML est constitué de balises (<balise>) qui comportent éventuellement un attribut (<balise attribut="valeur">). Ces balises permettent de structurer un contenu textuel, en en délimitant des parties à l'aide d'une balise ouvrante (au début de la partie) et d'une balise fermante (à la fin de la partie) : <balise> contenu </balise> (la barre oblique indique qu'il s'agit d'une balise fermante). Un fichier XML correspond donc à une structure arborescente.

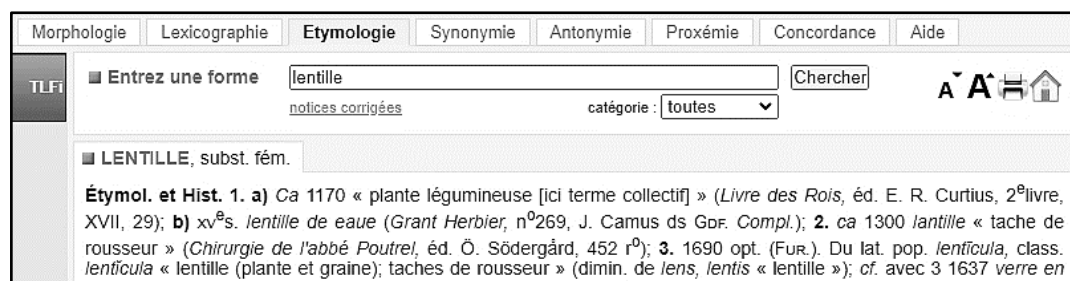


Figure 5. L'entrée H-É de l'article LENTILLE du TLFi affichée sur le site du CNRTL

Dans la suite de cette section, nous allons nous référer à l'entrée *lentille* pour présenter le contenu des fichiers XML. La figure 5 présente la notice H-É de *lentille* telle que visible sur le site du CNRTL. La figure 6 présente le code XML correspondant, et la figure 7 présente la macro-structure de l'arborescence associée. Cette structure est la même pour toutes les entrées : la balise <entry> délimite chaque entrée et comporte 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.

■ LA M  
Les qua

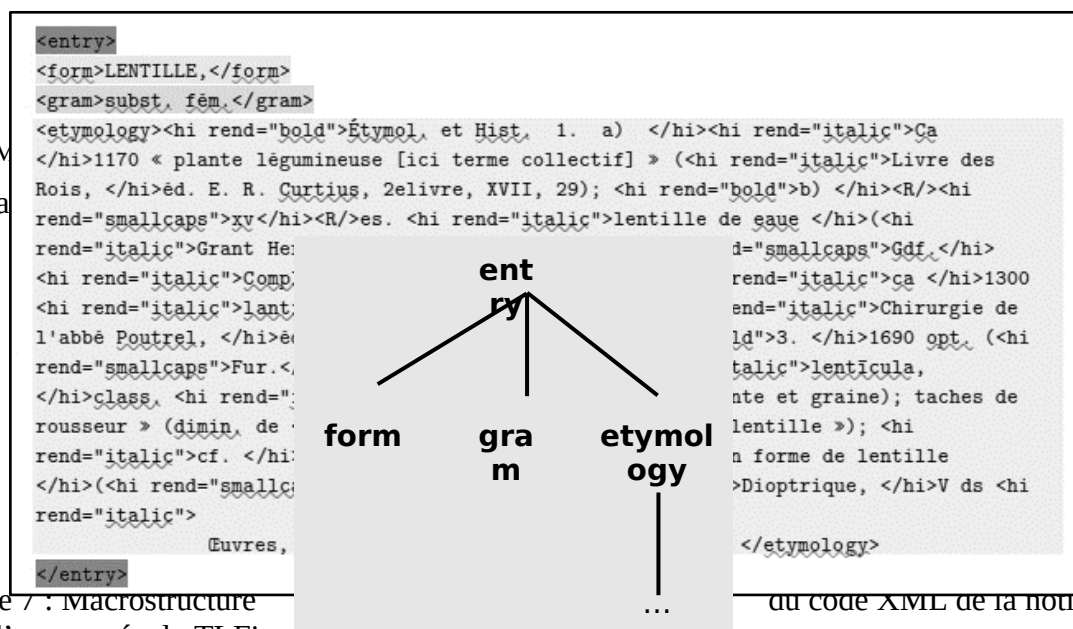


Figure 7 : Macrostructure H-É d'une entrée du TLFi

Usuellement, sur le web tout comme en XML, le fond et la forme d'un texte sont séparés, et une balise XML devrait toujours fournir une information fonctionnelle. Cependant, à l'exception de ces quatre balises, la très grande majorité des balises du TLFi ne sont pas fonctionnelles. Nous nous retrouvons alors avec des balises portant uniquement sur la forme, et c'est ce qui empêche de récupérer simplement les informations nécessaires à la visualisation de ces notices. Il s'agira donc d'exploiter les régularités syntaxiques et de mise en forme des notices pour retrouver les fonctions des différentes portions de texte identifiées.

Au sein de la balise <etymology>, il n'y a plus de structure arborescente, ou plus exactement, la structure est complètement plate, ce qui n'apporte pas d'information. Dans l'intégralité des fichiers XML, nous avons détecté 13 balises différentes à l'intérieur de la balise <etymology> : 3 sont dédiées à la création de tableaux, 1 à l'intégration d'images, 2 à la mise en exposant ou en indice de caractères, 3 balises sont des résidus d'une version plus ancienne des fichiers, et sont très peu présentes (24 occurrences au total dans l'ensemble des fichiers), 2 balises sont des balises fonctionnelles : la balise <var> ... </var> qui indique une variation et

la balise <emp> ... </emp> qui indique un emprunt. La dernière balise est la balise <hi>, qui indique une mise en forme (gras, italique, petites capitales) et qui est exploitée comme décrit dans la section 4.

L'algorithme déployé consiste à parcourir l'entrée, en recherchant des expressions régulières dans le texte, c'est-à-dire une portion de texte qui correspond à un motif donné (par exemple "trois ou quatre chiffres à la suite" pour désigner une année). Différentes expressions régulières sont utilisées pour identifier le siècle, les identifiants de niveaux et les définitions. Les collocations sont extraites en identifiant les portions de texte en italique.

L'algorithme général est le suivant :

```
Pour chaque balise <entry>
| récupérer son libellé entre les balises <form></form>
| récupérer la catégorie grammaticale entre <gram></gram>
| pour chaque niveau, rubrique et liste des mots italiques de cette entrée:
|   extraire le siècle
|   extraire les collocations, la définition et la sous-rubrique retenue
|   extraire la partie histoire de la partie étymologie à partir du contenu de la
|   rubrique
|   ne garder que les collocations qui sont dans la partie histoire
| remplir une ligne du tableau final avec
|   le libellé de l'entrée,
|   la catégorie grammaticale,
|   le niveau décelé,
|   le siècle,
|   la forme et ses collocations,
|   la définition,
|   la partie histoire
```

Jacques FRANÇOIS

CRISCO, EA 4255

Université de Caen- Normandie

Laurette CHARDON

CRISCO, EA 4255

Université de Caen- Normandie

Justine REYNAUD

GREYC, UMR CNRS 6072

Université de Caen- Normandie

## 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 / PIERRE 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. [cmlf2010\\_000268/cmlf2010\\_000268.html](https://cmlf2010_000268/cmlf2010_000268.html)>
- DARMESTER Antoine (1887) *La Vie des mots étudiée dans leurs significations*, Paris, Delagrave [rééd. Paris, Champ Libre, 1979]
- 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. Schuerer & 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 (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 2<sup>e</sup> Congrès Mondial de Linguistique Française*. La Nouvelle-Orléans, juillet 2010 <<https://www.linguistiquefrancaise.org/articles/cmlf/abs/2010/01/>>
- 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.
- GÉVAUDAN, Paul / WIEBEL, Dirk (2004), "Dynamic lexicographic data modelling. A diachronic dictionary development report", in: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* : 263-6.
- 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 (<sup>2</sup>1921|<sup>1</sup>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 / Philadelphia : Benjamins.
- 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  
WIERBICKA Anna (1996), *Semantics : Primes and universals*. Oxford / New-York : Oxford University Press.

## ■ Dictionnaires et base de données textuelle

ACADÉMIE FRANÇAISE (1694), *Dictionnaire de l'Académie Française*, 2 vol., Paris, J.-B. Coignard, 1694. Éditions consultables sur le site *Dictionnaire d'autrefois* de l'université de Chicago <<http://artflx.uchicago.edu/cgi-bin/dicos/pubdico1look.pl?>> : 1<sup>ère</sup> (1694), 2<sup>ème</sup> (1762), 3<sup>ème</sup> (1798), 8<sup>ème</sup> (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 IX<sup>e</sup> au XV<sup>e</sup> siècle*. Paris : Vieweg

HATZFELD Antoine / DARMESTETER Antoine (1890), *Dictionnaire général de la langue française*. Paris : Delagrave. <<https://gallica.bnf.fr/ark:/12148/bpt6k206410m.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 XIX<sup>e</sup> et du XX<sup>e</sup> siècle (1789–1960)*, 16 vol., 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], 25 vol., Bonn/Heidelberg/Leipzig-Berlin/Bâle, Klopp/Winter/Teubner/Zbinden, 1922–2002.

## Résumé en français

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<sup>e</sup> 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 2020, 2021, 2021b) 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  $\pm 29\,000$  entrées polysémiques (sur  $\pm 49\,000$  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 XII<sup>e</sup> siècle à nos jours. La méthode d'interprétation fonctionnelle des segments est explicitée en Annexe.

### Abstract in English

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 2020, 2021, 2021b), 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  $\pm 29,000$  polysemous entries (out of  $\pm 49,000$  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. The method of functional interpretation of the segments is explained in the Appendix.

### Zusammenfassung auf Deutsch

Diese Untersuchung widmet sich der Erforschung des Begriffs der "evolutiven Polysemie", der nicht auf einen bestimmten Typ lexikalischer Polysemie hinweist, sondern auf die Entwicklung des Spektrums der von einer Vokabel vermittelten Bedeutungen im Laufe der Jahrhunderte, hauptsächlich durch Bedeutungserweiterungen, da sich der Rückgang eines Teils von ihnen nur schwer dokumentieren lässt. Die allmähliche Anhäufung der Bedeutungen polysemer Vokabeln im zeitgenössischen Französisch ist seit mehr als einem Jahrhundert Gegenstand eingehender Studien, von den Untersuchungen A. Darmesteters (1887) über das FEW von W. von Wartburg (seit 1929) und die Einträge "Étymologie et Histoire" im Trésor de la Langue Française, die in den letzten drei Jahrzehnten des 20. Jahrhunderts auf dieser Grundlage verfasst wurden, bis hin zum historischen Wörterbuch von A. Rey (1992). Die Informatisierung des TLF hat zu zwei Ebenen der Bearbeitung geführt: formal und funktional (die verschiedenen Schriftarten und ihre kontextuelle Funktion) für die Haupteinträge, aber nur formal für die historisch-etymologischen Einträge, was eine Querschnittsuntersuchung verbietet. Unsere Studie, die sich auf drei frühere Veröffentlichungen (François 2020, 2021, 2021b) stützt, hat zwei Ziele. Zunächst geht es darum, die formale Behandlung dieser Einträge zu vervollständigen, indem wir verschiedenen Segmenttypen eine diachrone lexikografische Funktion zuweisen, insbesondere den Kennzeichnungen der zahlreichen Einträge (bis zu 146 für das Verb *tenir*), der Datierung des ersten Nachweises jeder Verwendungsart, den beteiligten Kollokationen, den Definitionen usw.).

Anschließend werden diese Datenklassen in einer grafischen Umwandlungstabelle zusammengefasst, die es ermöglicht, einen historischen Baumgraphen für die  $\pm 29.000$  polysemen Einträge (von insgesamt  $\pm 49.000$ ) automatisch zu erstellen. Der Zweck dieser grafischen Modellierung ist es, den Vergleich zwischen verschiedenen Arten von historischen Datenprofilen zu ermöglichen, die im letzten Abschnitt des Artikels in Kombination mit der diachronen Auswertung von FRANTEXT vom 12. Jahrhundert bis heute veranschaulicht werden. Die Methode zur funktionalen Interpretation der Segmente wird im Anhang erläutert.