

Evelyne Jacquey
Laurence Kister
Nancy-Université – atilf (UMR7118)
Evelyne.Jacquey@atilf.fr et Laurence.Kister@univ-nancy2.fr

Comparaison des structures thématiques de textes spécialisés et de thésaurus ou de terminologies

Introduction

Les travaux que nous menons proposent de déterminer si les concepts servant à structurer un domaine de spécialité via les termes du domaine sont utilisés pour structurer les textes de ce même domaine. Si c'est le cas, une seconde étape consistera à systématiser et modéliser la manière dont cette structuration s'opère. Ces travaux portent sur l'existence de liens entre la structure du discours scientifique et technique (mention des concepts) et la hiérarchie établie par le thésaurus sachant que celui-ci s'apparente à une représentation des connaissances dans un domaine fini à l'aide d'un langage de spécialité au même titre que la terminologie.

Cette approche s'articule autour de plusieurs types de questionnements qui relèvent de champs scientifiques différents. Du point de vue de la terminologie, la question qui émerge est celle de l'accès à une terminologie ou un thésaurus du domaine de spécialité. Pour la linguistique, nous disposons au laboratoire d'un thésaurus de la linguistique réalisé par le centre de documentation et de textes fondamentaux en linguistique dans la base Frantext, mais nous ne disposons pas de terminologie du domaine. Les questions qui se posent sont celles de la reconnaissance des termes du thésaurus dans les textes et du repérage de la structure thématique des textes. Pour effectuer ce repérage, nous nous appuyons sur l'hypothèse que les chaînes de référence contribuent à la structure thématique d'un texte. Cette hypothèse se traduit dans le cas de nos travaux par le repérage des chaînes de référence dans les textes. Afin d'explorer ces différentes questions, nous avons réalisé l'expérience que décrit cet article. Comme nous le verrons, cette expérience a confirmé certaines de nos hypothèses concernant la structuration thématique des textes par rapport à la structure terminologique des termes, mais nous a aussi permis de jeter des bases pour approfondir et détailler notre méthodologie actuelle.

1. Lien entre terminologie, thésaurus et sémantique lexicale

Un rappel des différents champs scientifiques dans lesquels s'inscrivent les travaux présentés dans cet article nous paraît nécessaire. En effet, deux champs scientifiques sont concernés par la confrontation de la structure terminologique des termes d'un domaine de spécialité et de la structure thématique des textes dans lesquels ces termes sont employés. Nos travaux s'inscrivent dans le domaine de la terminologie car ils reposent sur l'exploitation de thésaurus ou de terminologies. La sémantique lexicale tout comme la terminologie et le thésaurus s'intéressent à la description du contenu sémantique des items du lexique d'une langue, langue de spécialité en terminologie ou thésaurus. Certaines distinctions propres à la

sémantique lexicale permettent d'établir des corrélations entre la distinction générique/spécifique de la terminologie et du thésaurus et la distinction abstrait/concret, utilisée en sémantique lexicale.

La terminologie est une discipline scientifique qui étudie les vocabulaires spécialisés et qui en analyse les conditions d'utilisation. Elle est parfois présentée comme l'étude des vocabulaires de spécialité donc des termes propres à des domaines de connaissance délimités. Les termes qu'elle fait émerger peuvent appartenir tant à la langue courante qu'à la langue d'un domaine de spécialité. De ce fait, les termes peuvent faire l'objet tant d'une entrée dans un dictionnaire de langue que dans une base de données ou une banque de données terminologiques. Lorsque la terminologie se veut pratique, elle est souvent perçue comme une discipline au service d'autres disciplines : traduction, enseignement des langues étrangères, normalisation, rédaction scientifique et technique, etc. Elle partage avec l'ontologie la notion de concept : un terme est l'association d'une désignation (entité linguistique) et d'un concept qui en exprime la signification. La terminologie prend la forme de "fiches" contrairement au thésaurus et à l'ontologie qui sont représentés par des arborescences. En terminologie, les concepts font l'objet d'une description systématique qui regroupe différents types d'informations tels : le terme utilisé pour désigner le concept, la définition du concept via celle du terme utilisé, le domaine et les sous-domaines d'emploi du terme, les éventuelles variantes (orthographiques ou synonymes de type géographique, par exemple) et les équivalents dans d'autres langues. Parfois, une note et des exemples d'emploi apportent des informations complémentaires.

Le thésaurus parfois défini comme un lexique hiérarchisé ou un vocabulaire normalisé propre à un domaine n'a pas pour fonction de définir les termes qu'il organise. C'est un outil documentaire d'indexation (un document secondaire) qui permet de faire une description du contenu d'un document en texte intégral (document primaire). Le thésaurus est un ensemble structuré de termes destiné à faciliter la description d'un domaine, à harmoniser la diffusion, la gestion et le traitement de l'information propres à un domaine. Les termes ou *descripteurs* doivent être le moins polysémique possible d'où la nécessité d'écarter au maximum la synonymie et la quasi-synonymie. Le thésaurus est en quelque sorte un vocabulaire contrôlé : il dresse un inventaire des termes et les hiérarchise pour rendre compte de l'organisation du domaine. Cette organisation se matérialise par la présence de termes de différentes natures. D'une part, les génériques désignent les concepts principaux tandis que les spécifiques précisent les concepts particuliers propres à un concept principal. D'autre part, les équivalents correspondent à des variantes de certains spécifiques dans certains domaines ou sous-domaines et permettent de tenir compte de l'usage sur le terrain et assurer l'utilisation des spécifiques du thésaurus. Enfin, les associés rendent compte des liens de causalité, de localisation, de temps, etc. et servent à affiner la description ou à préciser les requêtes lors de la recherche d'information.

Malgré les différences qui existent entre le thésaurus et la terminologie ces deux types de représentation du lexique et des connaissances se rejoignent. L'organisation en générique et spécifique du thésaurus et à rapprocher de celle en terme, sous-domaine et domaine de la terminologie. Le fait que nous ne disposions pas, à l'heure actuelle, d'une terminologie mais seulement d'un thésaurus n'est pas bloquant pour les étapes de tests que nous avons entrepris avant d'élargir notre démarche à un corpus plus conséquent. Cette première étape exploite la distinction générique/spécifique du thésaurus et la confronte à la structure thématique des textes dans lesquels les termes apparaissent.

Tout comme la terminologie et le thésaurus, la sémantique lexicale s'intéresse via le sens des unités lexicales à la représentation des connaissances dans différents domaines.

En sémantique lexicale, le contenu sémantique des lexèmes peut être traduit grâce à des traits sémantiques usuels ([+animé], [+inanimé], [+humain], [+abstrait], [+concret], etc.). Le lexème *transcription*, par exemple, peut être interprété de différentes manières :

- [+processus] quand il traduit l'action de transcrire
- [+résultat] quand exprime ce qu'on obtient à l'issue du processus de transcription

Du point de vue de la distinction abstrait/concret, le substantif *transcription* est porteur du trait [+abstrait] dans le premier cas et du trait [+concret], dans le second cas.

La sémantique lexicale permet aussi l'analyse des relations sémantiques entre les lexèmes d'un lexique d'une langue : méronymie, synonymie, hyperonymie, hyponymie qui servent de fondement à la hiérarchie du thésaurus et à l'organisation en domaines et sous-domaines en terminologie.

Ces brefs rappels font émerger l'intérêt commun de la terminologie, du thésaurus et de la sémantique lexicale pour la description des concepts via les unités lexicales qui les désignent.

2. Chaînes de références et identification de la structure thématique du texte intégral

Les chaînes de références contribuent à l'identification des formes que prennent les désignations des concepts dans les textes. Parmi les différentes formes possibles, nous nous attachons tout particulièrement à l'examen des formes nominales correspondant aux termes, au sens où l'entend la terminologie : un terme désigne un et un seul concept dans un domaine déterminé (Cabré 1998 ; Béjoint & Thoiron, 2000). La confrontation de ces formes et de leur organisation dans les textes avec la structure du thésaurus vise à identifier en quoi la lexicalisation des concepts dans les textes s'appuie sur la distinction générique-spécifique du thésaurus. Cette confrontation repose sur l'hypothèse que plus un terme est à un niveau élevé de la hiérarchie dans le thésaurus, plus il est générique, phénomène qui a des répercussions pertinentes du point de vue de la sémantique lexicale sur les formes nominales correspondantes : plus le terme est générique plus il est abstrait du point de vue de la distinction [+concret]/[+abstrait] en sémantique lexicale. La mise en oeuvre de cette confrontation passe par l'annotation des formes nominales en [+concret] vs [+abstrait] : nous posons l'hypothèse selon laquelle les formes nominales annotées [+abstrait] ont une forte probabilité de correspondre à des termes plutôt génériques du thésaurus. L'annotation repose sur l'exploitation d'un lexique d'emploi de substantifs étiquetés [+concret] issu d'une acquisition automatique à partir du Trésor de la Langue Française informatisé (TLFi). Par contraste, nous faisons l'hypothèse que tous les substantifs qui ne figurent pas dans ce lexique ont des emplois préférentiellement [+abstrait] (Kister et Jacquey, 2006 ; Kister et Jacquey 2007a).

La confrontation des unités linguistiques annotées dans les textes et des termes issus des terminologies et/ou des thésaurus vise ainsi à révéler les similitudes et les distorsions entre l'ordre d'apparition dans le texte et le thésaurus. A moyen terme, nous cherchons à déterminer si les textes sont structurés selon une logique similaire à celle des thésaurus et si les points communs ou les divergences dans les deux types de documents facilitent ou entravent les activités d'analyse et d'indexation.

Afin d'illustrer le phénomène des chaînes de références (Boudreau & Kittredge, 2005 ; Kleiber, Schnedecker & Tyvaert 1997), nous proposons l'exemple suivant qui comporte une chaîne de référence autour de la *phonétique* extraite du *Cours de linguistique générale* de Saussure.

La physiologie des sons est souvent appelée " **phonétique** ". **Ce terme** nous semble impropre ; nous **le** remplaçons par **celui de** phonologie. Car **phonétique** a d'abord désigné et doit continuer à désigner **l'étude des évolutions des sons** ; l'on ne saurait confondre sous un même nom **deux études** absolument distinctes. **La phonétique** est **une science historique** ; **elle** analyse des événements, des transformations et **se** meut dans le temps. La phonologie est en dehors du temps, puisque le mécanisme de l'articulation reste toujours semblable à lui-même. Mais non seulement **ces deux études** ne se confondent pas, **elles** ne peuvent même pas s'opposer. **La première** est **une des parties essentielles de la science de la langue**.

Le concept introduit par *la physiologie des sons* devient successivement *phonétique*, *ce terme*, *le*, *celui de*, *phonétique*, *l'étude des évolutions des sons*, *deux études*, *la phonétique*, *une science historique*, *elle*, *se*, *ces deux études*, *elles*, *la première*, *une des parties essentielles de la science de la langue*. Cette chaîne de référence complexe contient des termes (*phonétique*), des syntagmes ne contenant pas de termes (*une des sciences historiques*) et des unités grammaticales (sujet : *elle*, *elles* – objet : *le*). Elle révèle aussi une autre forme de références, l'inclusion dans un ensemble avec *deux études*, *ces études*, *elles*. Nous pouvons aussi constater qu'un élément d'une chaîne peut être le point de départ d'une autre chaîne : *ce terme* (le terme *phonétique*) qui devient *celui de phonologie*. La *phonétique* et la *phonologie* quant à elle correspondent aux *deux études*.

A plus long terme, nous cherchons à déterminer si l'utilisation d'outils structurés tels terminologies, de thésaurus ou d'ontologies peut servir de fondement à une annotation sémantique des chaînes de références développées dans des textes scientifiques et techniques.

En conclusion de cet aperçu, nous résumons ci-dessous l'ensemble des hypothèses qui ont guidé l'expérience décrite dans la section suivante.

La confrontation de l'organisation des formes nominales dans les textes avec la structure du thésaurus que nous proposons consiste à identifier en quoi la lexicalisation des concepts présents dans les textes s'appuie sur la distinction générique-spécifique. Pour vérifier cette opposition nous formulons les hypothèses suivantes :

- un terme est générique s'il se situe à un niveau de hiérarchie élevé dans thésaurus ou s'il correspond à un domaine ou sous-domaine en terminologie,
- une forme nominale considérée comme abstraite en sémantique lexicale correspond à terme générique.

Pour évaluer la validité de ces hypothèses nous nous trouvons confrontées à la nécessité d'annoter les micro-thésaurus : repérage des termes, croisement avec la distinction de sémantique lexicale abstrait/concret.

3. Tests à partir d'une sélection de textes et de micro-thésaurus

Les tests que nous avons entrepris utilisent trois documents catégorisés de la base Frantext : *Cours de linguistique générale* (Saussure, 1968), *Le langage et la vie* (Bailly, 1925) et *La linguistique* (Perrot, 1957). Ces trois textes constituent un corpus de 105205 mots.

Pour ce qui est des documents secondaires, les tests ont été effectués avec quatre micro-thésaurus de Thésaulangue, le thésaurus mis au point et utilisé par le centre de documentation de l'ATILF. Il s'agit des micro-thésaurus *phonétique*, *syntaxe*, *sémantique* et *morphologie* qui d'un point de vue terminologique correspondent à quatre sous-domaines de la linguistique. Pour mener à bien cette phase de test, nous utilisons aussi une liste de candidats termes repérés lors de la lecture des textes intégraux. Ces candidats termes devraient permettre d'affiner les micro-thésaurus.

3.1. Répartition génériques/spécifiques

Les intitulés des quatre micro-thésaurus correspondent à des termes de premier niveau donc à des génériques. Les niveaux suivants proposent des termes de plus en plus spécifiques et ceux qui sont le plus droite de l'arborescence pour chaque domaine et sous-domaines sont les spécifiques. Le tableau ci-dessous présente le nombre d'occurrences des termes présents dans

les micro-thésaurus quel que soit leur niveau de généralité ou de spécificité et le nombre total d'occurrences de ces termes dans les documents en texte intégral.

	Nombre de termes	Nombre d'occurrences des termes
Phonétique	22	557
Syntaxe	59	501
Sémantique	13	179
Morphologie	13	149
Total	107	1384

Tab. 1 – Nombre de termes et nombres d'occurrences de ces termes pour les quatre micro-thésaurus

Le tableau ci-dessous permet de montrer que près de la moitié des termes du thésaurus sont présents dans les textes intégraux. Si on prend l'exemple de la phonétique, 22 termes du micro-thésaurus produisent 557 occurrences. Pour l'ensemble des quatre micro-thésaurus sélectionnés pour effectuer les tests, nous obtenons 1384 occurrences pour 107 termes représentés.

Si on détaille le nombre de termes en fonction de leur apparition aux différents niveaux de hiérarchie des micro-thésaurus, on obtient la répartition proposée dans le tableau présenté ci-dessous.

	Niveau 1	Niveau 2	Niveau 3	Niveau 4	Niveau 5	Niveau 6
Phonétique	26	31	403	95	Pas de terme dans le thésaurus	Pas de terme dans le thésaurus
Syntaxe	25	136	164	83	93	0
Sémantique	4	112	62	7	Pas de terme dans le thésaurus	Pas de terme dans le thésaurus
Morphologie	16	0	22	42	67	2
Total par niveau	71	279	651	227	160	2

Tab. 2 – Nombre d'occurrences de termes en fonction du niveau dans l'arborescence du thésaurus

Ce tableau montre que, pour les micro-thésaurus utilisés pour les tests, les termes les plus représentés ne sont ni les plus généraux, ni les plus spécifiques mais ceux qui se situent aux niveaux intermédiaires de l'arborescence qui représente le thésaurus. Afin de illustrer ce phénomène, nous proposons le détail pour le micro-thésaurus de la phonétique : un comptage a été effectué pour chacun des termes du micro-thésaurus représentés dans les documents en texte intégral.

			Saussure	Bailly	Perrot	TOTAL
Phonétique			16	1	9	26
	transcription		1			1
	phonétique descriptive					
		Son	206	13	11	230
		consonnes	55	2	2	59
		voyelles	48			48
		semi-voyelles	2	1	1	4
		durée	2	3		5
		liaisons	2	1		3
		mouillures				
		prosodie				
		accentuation	1			1
		intonation	1	2		3
		sonorité	8		1	9
		syllabe	52	1		53
		timbre	3	1		4
	phonétique auditive					
	phonétique expérimentale				1	1
		phonétique acoustique			1	1
		phonétique articulaire				
	phonétique historique					
		affaiblissement				
		amuïssement				
		assimilation	5	1		6
		consonnification				
		dégémination				
		dénasalisation				
		diphthongaison				
		dissimilation				
		épenthèse			1	1
		labialisation				
		métathèse				
		monophthongaison	3			3
		nasalisation	1			1
		palatalisation				
		sonorisation				
		vélarisation				
		vocalisation		1		1
	phonétique normative					
		orthoépie				
		orthophonie				
	phonologie		20	2	7	29
		phonème	80		6	86
		opposition phonologique				
		phonologie historique				

Tab. 3 – Nombre d’occurrences des termes de la *Phonétique* en fonction des documents et des niveaux de hiérarchie

3.2. Annotation sémantique en [+concret]/[+abstrait]

A l'issue de l'annotation automatique, on observe que parmi les 104 termes de Thésaulangue présents dans le lexique d'emploi [+concret], 53 termes sont atteints par la procédure automatique à l'aide de lexèmes tels *étude, science, action* [+abstrait] et 36 termes le sont à l'aide de lexèmes tels *mot, élément, catégorie* [+concret]. Cette première expérimentation de la procédure d'annotation ne prend pas en compte les termes complexes du thésaurus telles les constructions nominales, les locutions prépositives.

Afin de vérifier la validité des résultats obtenus, nous avons effectué une seconde annotation manuellement supervisée en repérant les lexèmes servant de descripteurs. Les résultats sont satisfaisants à quelques exceptions près. Si on prend l'exemple de la *phonétique*, les descripteurs représentés sont *perceptible, résultat, qualité, processus*. Les termes définis au moyen de *perceptible* et de *résultat* sont codés [+concret] ceux atteints par *qualité* et *processus* sont codés [+abstrait].

N1	N2	N3	N4		
phonétique				discipline	abstrait
	Transcription			processus	abstrait
			consonnes	résultat	concret
			voyelles	perceptible	concret
		liaisons		résultat	concret
		mouillures		résultat	concret
			accentuation	résultat	concret
		sonorité		perceptible	concret
		syllabe		syllabe	concret
		timbre		perceptible	concret
		affaiblissement		résultat	concret
		amuïssement		résultat	concret
		assimilation		résultat	concret
		dénasalisation		résultat	concret
		diphthongaison		résultat	concret
		dissimilation		résultat	concret
		épenthèse		résultat	concret
		labialisation		résultat	concret
		monophthongaison		résultat	concret
		nasalisation		résultat	concret
		palatalisation		résultat	concret
		sonorisation		résultat	concret
		vélarisation		résultat	concret
		orthoépïe		qualité	abstrait
		orthophonie		qualité	abstrait

Tab. 4 – Annotation supervisée manuellement

L'annotation semi-automatique ainsi réalisée doit, à terme servir à annoter les chaînes de références en texte intégral. Ces chaînes de références se matérialisent par des éléments lexicaux mais aussi des relations tels les anaphores (Amsili, Denis & Roussarie, 2005 ; Kleiber, 1994 ; Schnedecker, 1998). Le paragraphe suivant s'intéresse à la composition des chaînes de références que nous souhaitons annoter sémantiquement.

3.3. Observation des chaînes de références

Les observations que nous avons effectuées sur les chaînes de références montrent qu'elles se composent de termes, d'autres unités lexicales et d'unités syntaxiques. Le tableau ci-dessous

propose le détail des éléments qui composent les chaînes pour l'exemple du micro-thésaurus de la phonétique.

Type d'unité linguistique	%
Même déterminant + terme	5,5
Autre déterminant + terme	5
Autre déterminant + terme ou substantif	58
Déterminant + autre(s)/même(s)	2
Pronom sujet	10
Pronom objet	3
Pronom relatif	10
Autre relatif	3
Autres	3,5

Tab. 5 – Eléments linguistiques représentés dans les chaînes de références de la *phonétique*

Le repérage des chaînes de références pour la *phonétique* nécessite le comptage des unités linguistiques en fonction des différents types de références : lexicales, pronominales, relatives, etc. Pour les entités linguistiques, nous avons différencié les termes déjà présents dans le micro-thésaurus, des candidats termes repérés lors de la lecture et des autres unités linguistiques (substantifs quasi-synonymes, pronoms, etc.). Le tableau présenté ci-dessus montre que 68,5% des éléments de la chaîne de référence prennent une forme lexicale et utilisent un terme de Thésaulangue ou un substantif tandis que 13% correspondent à des pronoms personnels. Les relatifs (pronoms et autres relatifs) représentent 13,5% des éléments de la chaîne de référence.

Un examen un peu plus détaillé des résultats permet de conclure que les termes du thésaurus peuvent reprendre :

- un terme de même niveau appartenant à une autre sous-branche de la représentation arborescente (*phonème* reprend *son*)
- un terme de niveau inférieur appartenant à une autre sous-branche de la représentation arborescente (*phonème* reprend *voyelle* qui est un fils de *son*)
- un terme de même niveau appartenant à une autre sous-branche de la représentation arborescente avec une variation en nombre (*phonèmes* reprend *syllabe*).

Ces quelques exemples illustrent le fait qu'il est nécessaire de prendre en compte des informations plus complexes que la simple opposition générique/spécifique. Il s'agit de relations de quasi-synonymie entre des termes de sous-domaines distincts ou de l'usage atypique qui est fait de certaines unités linguistiques dans des contextes particuliers.

Pour ce qui est des candidats termes du corpus¹, nous en avons repéré plusieurs qui alternent avec *phonèmes* : aucun de ces candidats n'est en tête de chaîne et tous reprennent *phonème*. C'est, par exemple, le cas pour 6 des 42 occurrences d'*articulation*, 2 des 14 occurrences d'*occlusives*, 2 des 9 occurrences de *fricative*.

¹ cf. Annexe – Les candidats termes repérés lors de la lecture des documents.

En guise de conclusion

Les premiers tests effectués sur un corpus réduit et une sélection de micro-thésaurus sont encourageants. Ils montrent que les termes de niveaux intermédiaires, c'est-à-dire des termes plutôt spécifiques, sont plus nombreux dans le texte intégral que les génériques. De plus, ce résultat révèle que les têtes des chaînes de référence sont souvent des termes du thésaurus.

Les tests nous permettent aussi de constater que la prise en compte des relations propres au thésaurus du type *employé pour, voir aussi*, etc. lors du repérage des termes dans le texte intégral sont indispensables. Des tests d'annotation non automatique en texte intégral sont nécessaires pour envisager l'examen des divergences d'emploi entre le thésaurus et le texte intégral : le co-texte peut apporter des informations indispensables à l'annotation que la seule annotation automatisée du thésaurus ne peut pas saisir.

Une autre facette que nous devons aborder est l'interprétation ou la saisie d'un type particulier d'emploi par l'indexeur. Ce point demande qu'on s'intéresse à l'influence que le texte intégral exerce sur l'indexeur et qu'on explique pourquoi l'indexation semble se faire préférentiellement à l'aide de termes intermédiaires : ni les plus génériques, ni les plus spécifiques.

Les véritables enjeux, à plus long terme, sont l'utilisation de thésaurus et/ou de terminologie sémantiquement annotés pour annoter le document en texte intégral et l'examen de l'apport que pourrait constituer l'utilisation d'ontologies dans la procédure d'annotation.

Références

- Amsili, P., Denis, P., & Roussarie, L. (2005). *Anaphores abstraites en français : représentation formelle*, in J. Busquets et d. Hardt (eds), *Modèles et algorithmes pour la résolution d'anaphores*, Traitement automatique des langues, vol.46, n°1, pp.15-39.
- Bailly, C. (1925). *La linguistique et la vie*, Genève: Droz, Société de publications romanes et françaises.
- Béjoint, H., & Thoiron, P. (2000), *Le sens en terminologie*, Lyon : Presses universitaires de Lyon.
- Boudreau S. & Kittredge R. (2005), *Résolution des anaphores et détermination des chaînes de coréférences : différences entre variétés de textes*, in J. Busquets et d. Hardt (eds), *Modèles et algorithmes pour la résolution d'anaphores*, Traitement automatique des langues, vol.46, n°1, pp.41-69.
- Cabré, M.T. (1998), *La terminologie : théorie, méthode et applications*, Paris: Armand Colin.
- Kister, L., & Jacquy, E. (2006), *Traits sémantiques et anaphores pronominales*, 4^{ème} Rencontres de sémantique et de pragmatique, Orléans, 13-15 juin 2006.
- Kister, L., & Jacquy, E. (2007a), *Acquisition lexicale sémantique à partir de données lexicographiques au service de la comparaison entre des structures thématiques de textes spécialisés et de thésaurus*, Terminologie: approches transdisciplinaires, Gatineau, Canada, 2-4 mai 2007. (en ligne, <http://www.uqo.ca/terminologie2007/>)
- Kister, L. & Jacquy, E. (2007b), *NdeN et acquisition d'informations lexicales à partir du Trésor de la Langue Française Informatisé*, in COgnition, Représentation, Langage, Vol.5, n°1, (en ligne, <http://edel.univ-poitiers.fr/corela/>).
- Kleiber, G. (1994), *Anaphores et pronoms*, Champs linguistique, Louvain-la-Neuve: Editions Duculot.
- Kleiber, G., Schnedecker, C., & Tyvaert, J.E. (1997). *La continuité référentielle*, Recherches linguistiques, Paris: Klincksieck, n°20.
- Perrot, J., (1957). *La linguistique*, Paris : PUF, Que sais-je ?
- Saussure, F., (1968). *Cours de linguistique générale*, Paris: Payot.
- Schnedecker, C. (1998). *Les corrélats anaphoriques*, Recherches linguistiques, Paris: Klincksieck, n°22.

Annexe

Les candidats termes repérés lors de la lecture des documents

	Saussure	Bailly	Perrot	TOTAL
langue	607	216	60	883
signe	160	33	8	201
articulation	42	2	2	46
système	147	19	32	198
image	26		4	30
phonation	14		2	16
altération	28		1	29
patois	7	9		16
chaîne	21			21
implosion	24			24
explosion	16			16
fermeture	12			12
ouverture	16			16
occlusive	14		1	15
fricative	9			9
nasale	11		1	12
liquide	7			7
aperture	39			39
sonante	13			13
diphthongue	13			13
signifiant	46	7	11	64
signifié	24	7	7	38
génitif	10	5		15
synchronie	13		4	17
état de	17	2	4	23
signification	33	5	5	43
unité	198	2	1	201
entité	29	1		30
flexion	27			27
désinence	29	3		32
agglutination	24	2		26
terme	90	3	13	106
stylistique		21	1	22
forme	230	34	6	270
expression	28	21	3	52