Traitement automatisé de la néologie : pourquoi et comment intégrer l'analyse thématique ?

Christophe Gérard, Ingrid Falk, Delphine Bernhard

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

Traitement automatisé de la néologie : pourquoi et comment intégrer l'analyse thématique ?

Gérard, Christophe, & Falk, Ingrid, & Bernhard, Delphine

LiLPa – Linguistique, Langues, Parole
EA 1339, Université de Strasbourg
{christophegerard,ifalk,dbernhard}@unistra.fr

1 Introduction

Cet article a pour cadre un projet en cours nommé *Logoscope* et développé à l'Université de Strasbourg (2012-2015). Le cœur de ce projet se constitue d'un programme informatique qui, chaque jour, scrute les pages web de la presse quotidienne francophone (*Le Monde*, *La Croix*, *L'Equipe*, *Dernières Nouvelles d'Alsace*, etc.) à la recherche de mots nouveaux, dans le but de produire une ressource *dynamique* (i.e. enrichie sans limites dans le temps) utile à différentes communautés d'utilisateurs.

L'originalité du *Logoscope* le distingue des outils similaires existants : l'acquisition semi-automatisée des néologismes se concentre sur les conditions textuelles et discursives de l'innovation lexicale. Cette direction de recherche est en effet essentielle : dans le domaine de la néologie, ce que nous savons déjà de la morphologie d'une langue et des types de formation des mots doit être complété par une description des conditions de production et de réception des néologismes, étant donné qu'aucune création lexicale ne se produit jamais en dehors d'un texte particulier et d'une situation de communication définie.

Pratiquement, nous proposons donc un outil d'acquisition et une ressource néologique qui documente non seulement les néologismes au moyen des variables traditionnelles (morphologie, formation du mot, parties du discours, etc.), mais aussi au moyen de variables chargées de décrire de manière précise le contexte communicationnel où ils apparaissent.

Si l'implémentation informatique de cette approche, dans une optique d'automatisation, rencontre à l'évidence nombre de difficultés, principalement pour l'identification des thèmes du texte et la catégorisation de son genre discursif (éditorial, interview, tribune libre, etc.), celles-ci sont toutefois loin d'être insurmontables. Cet article entend commencer de le montrer en exposant le cadre théorique, les bases techniques et les résultats actuels de notre classification thématique, celle-ci ayant pour effet inattendu d'aider le néographe à repérer les bons néologismes-candidats.

2 Pour une approche textuelle et discursive de la néologie

Les approches de la création lexicale se laissent regrouper en deux grands ensembles conceptuels : les approches qui relèvent d'une problématique du texte, et celles qui relèvent d'une problématique du signe, qui est la plus répandue et la plus ancienne. Cette dernière étudie traditionnellement, depuis le XIXe siècle, la formation des mots (de Darmesteter, 1888 à Sablayrolles et Pruvost, 2003) et leurs règles de construction

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20140801208

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

morphologiques (Kastovsky, 2006 ; Kerleroux, 2006). La problématique du signe fait ainsi de la langue un niveau d'analyse privilégié, voire exclusif.

Historiquement, la problématique du texte prend quant à elle son essor dans les années 1970. C'est la germanistique allemande (Peschel, 2002 : 67–85) qui approfondit l'idée que toute création lexicale ne fait sens que pour le texte qu'elle constitue et dans la situation de communication où elle est exprimée et interprétée. Ces travaux n'examinent pas seulement les fonctions textuelles (cohésion et progression) des néologismes (e.g. Dederding, 1983 ; Peschel, 2002 ; Gérard 2011), mais aussi les corrélations qui associent certains procédés de création lexicale à certains types de discours et à certains genres textuels (Matussek, 1994 ; Fleischer et Barz, 1995 ; Barz et alii, 2000 ; Siebold, 2000 et 2005 ; Cabré et alii, 2003 ; Elsen, 2004 ; Elsen et Dzikowicz 2005 ; Ollinger et Valette, 2010).

Tous ces travaux montrent que les conditions textuelles et discursives jouent un rôle crucial dans la compréhension du fonctionnement des néologismes. Ainsi, par exemple, la position du néologisme dans le texte (titre de l'article, corps du texte ou note de bas de page ; début ou fin de paragraphe) n'est jamais neutre. De même la thématique d'un texte est liée au(x) néologisme(s) qu'il accueille : tout néologisme s'inscrit nécessairement dans la progression thématique de son texte d'accueil. Ainsi, dans l'exemple suivant *loto-impôt* conjoint les thèmes du Jeu et de la Finance :

> Nous n'aimons pas les impôts, mais nous aimons les jeux de hasard. Alors pourquoi ne pas jouer au **loto-impôts** ? En payant 10 % de majoration on pourrait gagner le remboursement de l'impôt payé et une exonération d'impôt pendant 10 ans. (Événement du jeudi, 07/11/1985, courrier des lecteurs).

Du reste, surtout, la thématique est un critère nécessaire pour identifier l'innovation polysémique (ex. flûte [musique] ; flûte [alimentation], flûte [marine]) et un critère utile pour l'identification semi-automatisée des néologismes (voir aussi infra 5). Quant au genre textuel, autre variable centrale, il influence non seulement la quantité de création lexicale (la nécrologie n'est pas néologène) mais aussi la qualité de ces créations. Par exemple, Ollinger et Valette (2010) ont montré que la création lexicale varie selon le type de magazines (voir Tableau 1).

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

	Marianne		**Le Point**		**Le Nouvel Observateur**	
	Form.	**Exemples**	**Form.**	**Exemples**	**Form.**	**Exemples**
Opposition, Négation	18	*anticorporatiste*	6	*non-annonces*	4	*anticoncurrentiel*
Péremption	18	*ex-trublion*	4	*ex-candidate*	4	*ex-travailleuse*
Approximation	4	*quasi-maniaque*	0		1	*quasi-impasse*
Hyperbole	9	*hypercapitalisme*	2	*surprofit*	6	*superprofits*
Itération	10	*refondation*	3	*remobiliser*	0	
Agglutination	7	*tactico-politiciens*	2		0	
Procès (-iser, -isation)	7	*starisation*	3	*annualisation*	1	
Dérivation d'ent. nom.	26	*gaudinerie, Sarkozie*	1	*villepeniste*	9	*berlusconien*
Total	**99**		**21**		**25**	

Tableau 1 : Types de créations lexicales selon (Ollinger et Valette, 2010).

La réalité textuelle et discursive des créations lexicales n'est donc plus à démontrer, mais beaucoup reste à étudier, en particulier l'incidence des normes de genre sur la création lexicale. Mais seul le permettra la constitution des ressources linguistiques qui nous manquent aujourd'hui. Pour ce faire, en amont, il est primordial de se donner une conception claire des objectifs de description (section 3), puis de développer des outils aptes à réaliser ces objectifs (sections 4 et 5).

3 Veilleurs de néologie : état des variables descriptives

En néographie, la cohérence d'ensemble des variables descriptives comprend deux niveaux hiérarchisés. Au niveau supérieur se trouve la distinction entre les variables lexicales et contextuelles, au niveau inférieur celle entre les variables discursives et textuelles, soit schématiquement :

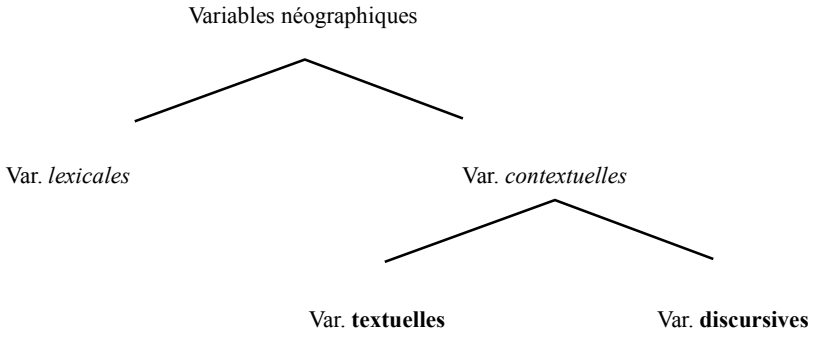

Figure 1 : architecture des variables néographiques.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

Nous entreprenons ici de fonder ces distinctions, utiles en général pour la néographie, pour situer notre projet relativement aux veilleurs de néologie existants, lesquels ne prennent en compte qu'une partie des variables contextuelles, celles-ci n'étant par ailleurs généralement pas documentées à l'aide de processus automatiques.

3.1 Variables lexicales

Plusieurs équipes de recherches européennes proposent aujourd'hui, pour différentes langues, des systèmes de détection semi-automatisée de néologismes. Comme elles ont pour objectif d'élaborer une forme de dictionnaire, leur tâche principale consiste à établir une *fiche d'attestation* pour chaque néologisme repéré. Cette fiche, aujourd'hui modernisée au format électronique, note des informations strictement lexicales, bien entendu, mais aussi des informations liées au contexte d'apparition du néologisme. Un examen comparé des fiches utilisées par certains veilleurs de néologie[1] révèle cependant des capacités de documentation néographique plus ou moins étendues, au sens où chaque veilleur comporte un nombre plus ou moins grand de variables descriptives.

Ainsi, concernant tout d'abord la documentation lexicale des néologismes, les variables retenues par la WORTWARTE (dédié à l'*allemand* : Lemnitzer, 2010)[2] apparaissent-elles très réduites en comparaison avec le veilleur OBNEO (dédié au *catalan* et au *castillan* : Cabré et alii, 2003 ; Estopà et Cabré 2004), ce dernier présentant à son tour moins de possibilités que NEOLOGIA (dédié au *français* : Cartier et Sablayrolles 2008 ; Sablayrolles 2011 ; Cartier 2011), dont l'interface web permet au néographe de détailler les grands secteurs suivants[3] :

- Morphosyntaxe : parties du discours / sous-catégorie grammaticale (ex. adj. qualificatif) / nombre / genre
- Sémantique : prédicat, argument ou actualisateur / hyper-classe (humain, animal, locatif, état, etc.) / domaine lexical[4]
- Néologie : types de formation (formels/sémantiques) / configuration morphologique / configuration phonologique / influence linguistique[5]
- Relations sémantiques : synonymie, antonymie ou hyperonymie

Si NEOLOGIA ne renseigne pas systématiquement les relations sémantiques, et si beaucoup de variables sont évidemment communes à d'autres veilleurs (partie du discours, type de néologisme, domaine lexical, etc.), la précision lexicale de cette base (ex. sous-catégorie grammaticale) et son intérêt particulier pour la néologie sémantique (Gérard et Kabatek, 2012) en font un outil d'analyse des plus complets. Malgré tout, et outre la variable qu'OBNEO nomme *variante* (orthographique, morphologique, syntaxique), une variable supplémentaire reste étonnamment ignorée par NEOLOGIA et par les autres veilleurs : aucun ne documente, au-delà de la *langue* du néologisme, la *variation linguistique* (diatopique, diastratique, diaphasique).

Or ce déficit sociolinguistique a pour conséquences de priver la néographie d'une dimension indissociable des phénomènes de création et de diffusion lexicales[6], alors que les dictionnaires de langue intègrent cette variable et que les mots détectés au hasard par les veilleurs témoignent constamment de styles de langue différents, par exemple. En outre, alors que pourraient émerger *des* français d'Afrique (Calvet 2010 : ch. 5), par exemple,

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

l'heure est à la constitution de bases variationnelles (Projet Varitext[7]). Reste le problème pratique du néographe : comment documenter la variation linguistique, voire comment l'automatiser ? À cet égard, la prise en compte du genre de discours (oral / écrit) s'avère un indicateur de confiance[8].

3.2 Variables contextuelles

Concernant justement la documentation du contexte, ces trois veilleurs de néologie connaissent également des points communs et des différences notables. Une variable déterminante partagée par les trois veilleurs précédents correspond au *cotexte* d'apparition du néologisme, documenté sous forme d'une phrase ou d'un paragraphe. L'importance du cotexte va sans dire puisqu'il sert tout à la fois de preuve d'existence de la création lexicale, d'exemple d'emploi (à la manière d'une citation de dictionnaire) et surtout d'interprétant pour comprendre le sens du néologisme, ce qui est essentiel pour l'utilisateur des fiches dès lors que le néologisme n'y est pas défini par le néographe. En plus du cotexte, les veilleurs complètent leur documentation par diverses variables :

- Date d'apparition : jour-mois-année
- Source : médialité (orale/écrite) / nom du média / page / domaine de compétence[9]
- Enonciation : nom de l'auteur, nom du transmetteur / type d'auteur, type de transmetteur, etc.[10]
- Typographie : tirets, guillemets, italiques, etc.

Mais le plus remarquable ici est la présence d'une variable qui singularise chaque veilleur par rapport à ses homologues : NEOLOGIA est le seul à détailler la *position* du néologisme dans le texte (corps du texte, titre, sous-titre, chapeau, légende, etc.), OBNEO attire notre attention sur les *genres de discours* et WORTWARTE propose une *classification thématique* semble-t-il inédite des néologismes. Ces trois variables pourraient paraître de second plan et par là même d'une utilité néographique toute relative. Leur examen, pourtant, permet d'une part de concevoir les variables contextuelles avec plus de cohérence et, d'autre part, de préciser la valeur néographique des dimensions textuelle et discursive de la néologie.

3.3 Variables textuelles et discursives

Inédite à notre connaissance, en néologie, la distinction entre variables textuelles et discursives éclaire la diversité des variables contextuelles. Les variables *discursives* sont ainsi nommées en référence à la notion de *tradition discursive* développée dans le champ de la romanistique allemande[11]. Cette notion, qui renvoie à des règles et à des normes d'expression transmises au sein d'une communauté (dans une période qui peut être limitée), recouvre aussi bien les genres conversationnels, les genres textuels, les anciens registres rhétoriques que les styles collectifs (ex. le burlesque au XVIIe siècle). Tout domaine de discours (science, religion, médecine, littérature, etc.) abrite ses propres traditions discursives. Dans le domaine journalistique, qui est la source privilégiée des veilleurs de néologie, les règles et normes expressives sont données par les *genres de discours* (écrits / oraux) et le *style collectif* de la publication/émission[12]. La base d'OBNEO a le mérite

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

d'organiser explicitement son sous-corpus oral, issu de médias télévisés et radiophoniques, en intégrant la variable discursive des *genres* (voir Tableau 2).

genres radiophoniques	**genres télévisés**
interview	interview
nouvelles	nouvelles
talk show	débat
…	magazines
	documentaires
	…

Tableau 2 : Genres du sous-corpus oral de la base d'OBNEO.

OBNEO se fait ainsi le miroir de la diversité des pratiques médiatiques (orales), une diversité dont le rôle est important dans ce domaine puisqu'elle sert à garantir en partie l'audience du journal, organe de presse non monocorde. Par ailleurs, comme les autres veilleurs de néologie, la base catalane et castillane renseigne sur le nom du média (ex. *Catalunya Ràdio*) ainsi que sur le nom donné à l'émission (*Cap a la Neu*, *Debat Jove*, *Versió Original*, etc.). Documenter ces titres comme une variable discursive, et pas simplement comme une dénomination utile, permet d'identifier des *instances d'énonciation* (collectives) dont on pourra ultérieurement étudier le style (collectif) en termes de quantité et de qualité de néologie (types de formation privilégiés, etc.) : au sein d'un média particulier (une radio, une chaîne de télévision, un journal) ou d'une émission radiophoniques / télévisée particulière.

Indissociablement liée à un texte individuel, la *position* du néologisme est une des variables *textuelles*, comme la page, la typographie, la médialité (orale/écrite), l'énonciation et la thématique du texte. Pour la néologie, la position textuelle est un indicateur de différences qualitatives : il est bien connu que les titres (dits incitatifs) conditionnent spécialement la production néologique ainsi que l'effet rhétorique produit sur le lecteur. Par ailleurs, les parties initiale et finale du corps de texte sont des zones rhétoriquement stratégiques où les néologismes gagnent à être observés (Loiseau, 2012).

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

3.4 Diversité de la variable thématique

Bien qu'elle soit une variable textuelle majeure, la *thématique* n'est jamais documentée comme telle par les veilleurs de néologie. Seule la WORTWARTE emploie ce terme comme descripteur, nous l'avons dit, pour classer les néologismes en grands domaines : automobile, biotechnologie, santé, société, musique, mode, économie, etc. Certes, du point de vue lexical, ces informations reviennent à noter le *domaine* du néologisme, comme le fait l'entrée d'un dictionnaire :

Burnoutsyndrom [santé]
Acrylamid [alimentation]
Slow-Photography [art]
Tempokrat [politique]
...

Mais, du point de vue contextuel, la classification thématique des néologismes a une conséquence pratique de taille : dans la WORTWARTE l'utilisateur dispose d'un accès thématique à la base de néologismes, en plus de l'accès traditionnel par entrée néologique. Le choix d'un thème, *santé* par exemple, donne la liste de tous les néologismes qui lui sont associés :

> Abnehm-Aktion, Acrylamidbelastung, adrelaninträchtig, Age-Scan-Computer, Age-Scan-Test, Algenmassage, Alltagsmotorik, Altersprävention, Amakaphobie, Anthraxerreger, Anti-Aging-Beratung, Anti-Aging-Branche, Anti-Aging-Institut, Anti-Aging-Produkt, Anti-Aging-Trend, Antibiotika-Mast, …

Cette thématisation de la ressource néologique, ou ce qu'on peut encore appeler la *conversion thématique* d'une variable lexicale — le domaine lexical du néologisme — doit se baser sur un répertoire conventionnel de noms de domaines assez puissant pour étiqueter n'importe quelle création lexicale inédite : le répertoire du Petit Robert semble remplir cette condition.

Mais la conversion thématique du domaine lexical n'est qu'une des trois manières de concevoir la documentation thématique des néologismes en situation journalistique. En effet, tout d'abord, l'utilisation de *rubriques* (politique, économie, culture, sport, etc.) dans la presse quotidienne traduit une réalité thématique certaine. Il s'agit d'une variable textuelle qu'on peut dire globale au sens où elle permet d'organiser l'information du journal dans son ensemble. Pour l'écrit, OBNEO documente cette variable en reprenant *verbatim* les noms de rubriques utilisés par chaque journal (Estopà et Cabré, 2004 : 26-30) : comportant jusqu'à 30 étiquettes, les répertoires de ces noms de rubrique ont le désavantage d'être relativement hétérogènes (ex. dans *El Temps*, on trouve : calendrier, sciences, éditorial, idées, etc.). Au contraire, pour l'oral, les nouvelles radiophoniques et télévisées sont documentées au moyen d'un répertoire réduit, mais plus homogène et transversal aux différentes chaînes médiatiques :

- sciences et technologie
- culture et spectacles
- économie

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

- sport
- météorologie
- politique (thèmes de politique nationale et internationale)
- publicité
- société (événements sociaux divers)
- transit (informations sur l'état des routes et du transit en général)

Enfin, la troisième manière de concevoir la documentation thématique des néologismes est la plus proche du texte. C'est une variable textuelle locale qui renseigne notamment sur le sujet particulier du texte où se trouve le néologisme. Elle fait précisément l'objet de l'analyse thématique automatisée.

4 Approches existantes de l'analyse thématique automatisée

4.1 Notion de thème en traitement automatique des langues

Les outils d'analyse thématique automatisée utilisent une représentation simple et pragmatique du thème sous forme d'une liste de mots caractéristiques. En effet, un mot seul est généralement ambigu mais, phénomène sémantique très ordinaire, l'association de plusieurs mots dans une liste permet de lever les ambiguïtés et, ainsi, de définir un thème. Ainsi entendue, l'analyse thématique peut viser divers objectifs, qu'on va parcourir dans cet ordre :

- Détection de thèmes dans un texte afin de le segmenter en sous-unités thématiquement homogènes
- Détection de thèmes dans un corpus textuel afin de le caractériser
- Annotation thématique de documents à partir de thèmes prédéfinis

4.2 Segmentation thématique

Une des applications les plus anciennes de l'analyse thématique est celle de la segmentation automatique de textes en fonction des ruptures thématiques. L'objectif est donc de repérer les changements de thème dans un document sans pour autant expliciter les thèmes présents dans un texte. L'algorithme TextTiling, proposé par Hearst (1997), repose sur l'idée que les ruptures thématiques correspondent à des modifications du vocabulaire utilisé. La première étape de la méthode consiste à découper les textes en mots et en séquences de mots d'une longueur donnée. Les blocs adjacents de séquences de mots (qui correspondent approximativement à un paragraphe) sont comparés de manière à obtenir un score. Les frontières sont identifiées aux positions où les scores sont les plus bas, indiquant la cohésion lexicale la plus faible.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

4.3 Détection des thèmes d'un corpus de textes

Les modèles thématiques probabilistes (*topic models*) permettent de détecter automatiquement les thèmes d'un corpus de textes, sans a priori et sans recours à une ressource externe au corpus. Différents modèles de ce type ont été proposés, notamment l'Allocation Dirichlet Latente (LDA : *Latent Dirichlet Allocation*) par (Blei et alii, 2003). Ces modèles identifient des thèmes « latents » dans un corpus textuel donné en entrée, sous forme de listes de mots caractérisant chaque thème. Un document est composé d'un nombre fini de thématiques et chaque thématique a un poids plus ou moins grand dans le document. De même, le corpus contient un nombre fini de thématiques d'importance variable. Les thématiques extraites ne sont toutefois pas nommées et un travail d'analyse est donc nécessaire pour leur attribuer une étiquette. C'est ce type d'approche que nous utilisons dans nos travaux.

4.4 Annotation thématique de documents

Lorsqu'une ressource thématique est disponible, celle-ci peut être utilisée pour l'annotation. Dans ce cas, la granularité de l'analyse thématique dépend de la granularité de la ressource utilisée. Beust (2002) propose un outil de marquage thématique appelé ThemeEditor (https://beust.users.greyc.fr/ThemeEd). L'outil utilise des listes de mots définies manuellement pour caractériser chaque thème ou isotopie. Dans le texte, une couleur est affectée à chaque thème afin de surligner les mots associés et ainsi les mettre en valeur.

Dans le domaine de l'indexation de textes, les mots-clés ou expressions clés caractérisent les thèmes principaux d'un document. Medelyan et Witten (2008) proposent une méthode d'indexation automatique qui utilise un thésaurus existant comme répertoire contrôlé de termes d'indexation. Ceci permet d'éviter les problèmes d'une extraction libre de mots-clés à partir de critères purement statistiques. L'outil proposé par Medelyan et Witten identifie tout d'abord les termes du thésaurus dans le texte, puis utilise un modèle obtenu par apprentissage supervisé pour sélectionner uniquement les termes les plus pertinents. La sélection repose sur différents types de propriétés : score tf.idf[13], position de la première occurrence dans le document, longueur du terme en mots, nombre de liens sémantiques avec d'autres candidats identifiés dans le document. Comme pour toutes les méthodes par apprentissage supervisées, il est nécessaire de fournir un ensemble de documents indexés manuellement afin de procéder à l'apprentissage du modèle. L'effort d'annotation reste toutefois limité et a été estimé entre 50 et 100 documents.

5 Expérimentations en cours

Cette section présente le développement en cours d'un outil de détection semi-automatique de néologismes. Le fonctionnement envisagé est le suivant : notre outil récupère chaque jour des articles de presse en français disponibles en ligne. Ces articles sont ensuite pré-traités pour récupérer le contenu textuel, ils sont découpés en phrases, puis en mots (signes). Ces mots sont ensuite filtrés par une liste d'exclusion. Elle est basée principalement sur Morphalou (Romary et alii, 2004) et les listes de mots de Wortschatz (Quasthoff et alii, 2006). Nous avons également appliqué l'outil CasEN (Maurel et alii, 2011) pour reconnaître et exclure des entités nommées. Un résultat typique d'un tel pré-traitement est montré en Figure 2.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

lmd (18)	twitter/widgets (7)	india-mahdavi (3)
pic(this (18)	garde-à (6)	kilomètresc (2)
lazy-retina (9)	ex-PPR (4)	geniculatus (2)
onload (9)	pro-Morsi (4)	margin-bottom (2)
onerror (9)	tuparkan (4)	politique» (2)
amp;euro (7)	candiudature (3)	. . .

Figure 2 : Les mots inconnus les plus fréquents, extraits des articles de presse du 2013-07-12. Entre parenthèses la fréquence.

Cet exemple montre clairement les problèmes de cette approche basée pourtant sur des méthodes conformes à l'état de l'art : la plupart de ces signes ne sont pas des néologismes-candidats intéressants et ne sont souvent même pas des mots possibles de la langue (*candiudature*, *pic(this*, *margin-bottom*, etc.).

Les expériences que nous présentons ici viennent appuyer deux hypothèses de travail. D'une part elles montrent qu'en utilisant des méthodes statistiques d'apprentissage automatique (supervisé) il est possible de trier les mots/signes inconnus d'une façon plus pertinente, c'est-à-dire plus appropriée à notre tâche. D'autre part, elles font apparaître l'impact et l'utilité des traits relatifs à la thématique textuelle pour l'identification des mots/signes inconnus.

5.1 Méthode d'apprentissage automatique

Pour nos expériences, notre outil a récupéré les articles de presse en français à partir des flux RSS des journaux suivants : *Le Monde* (659), *Libération* (504), l'*Équipe* (594), *Les Echos* (956) (en parenthèse le nombre d'articles récupérés), pendant 7 jours en juillet/août 2013. Les mots issus de ces articles ont ensuite été filtrés à l'aide de notre liste d'exclusion. Cette procédure a produit une liste de 692 mots/signes inconnus.

Pour identifier les néologismes-candidats les plus probables nous utilisons un classifieur automatique basé sur des exemples préalablement étiquetés et des traits associés à chaque item à classer (les mots/signes inconnus repérés). En ce qui concerne les exemples préalablement étiquetés, dans notre cas, un expert linguiste a estimé que 81 des 692 mots inconnus détectés étaient de bons candidats. Quant aux traits sur lesquels repose l'apprentissage, ils ont été extraits du corpus des textes collectionnés. Nous avons ainsi identifié trois types de traits : 1) traits relatifs à la forme du mot/signe, 2) traits morpho-lexicaux et 3) traits relatifs au contexte thématique textuel.

Les traits du premier groupe relèvent de la forme ou construction du mot/signe et sont indépendants de la langue (ex. nombre de caractères, présence ou non de majuscules ou de signe non-alphabétiques, fréquence). Les traits morpho-lexicaux dépendent de certaines caractéristiques du français (présence de préfixes ou suffixes particuliers, orthographe, etc.). Enfin, les traits du troisième groupe sont conçus pour permettre un

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

accès au contenu thématique textuel dans lequel apparaît le mot inconnu. Le cadre méthodologique et leur extraction seront détaillés en Section 5.2.

Sur la base de différentes combinaisons de ces trois groupes de traits, et l'estimation de l'expert, on constitue puis évalue différents modèles de classement (un classifieur pour chaque combinaison de traits). La performance de ces modèles permet de jauger l'apport des différents groupes de traits pour la détection semi-automatique de néologismes.

5.2 Traits contextuels thématiques

Notre outil se propose d'observer notamment la création lexicale dans un contexte textuel thématique plus large que la phrase ou le paragraphe. Dans les expériences présentées ici nous avons visé cet objectif en nous appuyant sur la technique (probabiliste) du *topic modeling* (voir Section 4.3).

Nos données étant issues de contenus journalistiques, nous tentons dans un premier temps d'obtenir une estimation des thématiques traitées dans l'ensemble de journaux que nous observons. Pour cela nous avons récupéré une collection de 4 755 textes des journaux suivants : *Le Monde* (898), *Libération* (269), *La Libre* (1784), *Presseurop* (690), *Le Journal de Dimanche* (206), *Rue89* (212), *Slate* (74), *L'Équipe* (892) - en parenthèses le nombre de textes pour chaque journal. Cependant nous considérons que ce ne sont que les mots d'un certain vocabulaire qui sont d'un apport significatif pour la détection des thèmes. Ce vocabulaire inclut les mots les plus fréquents du français (~ 10 000 mots) mais exclut les mots vides. Suivant cette logique, pour les textes collectés, nous ne gardons que les mots pleins de ce vocabulaire (~ 7 900 mots).

À partir de ces textes filtrés, nous utilisons l'outil Mallet (McCallum, 2002) pour inférer un ensemble de 10 thématiques. En Figure 3 nous présentons une visualisation des 10 thématiques découvertes sur ce corpus à l'aide de l'outil Termite (Chuang, Manning, & Heer, 2012). Cet outil permet de montrer, pour chaque thématique, les mots les plus saillants. Ainsi, dans cet exemple, la thématique 2 pourrait être étiquetée *Politique*.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

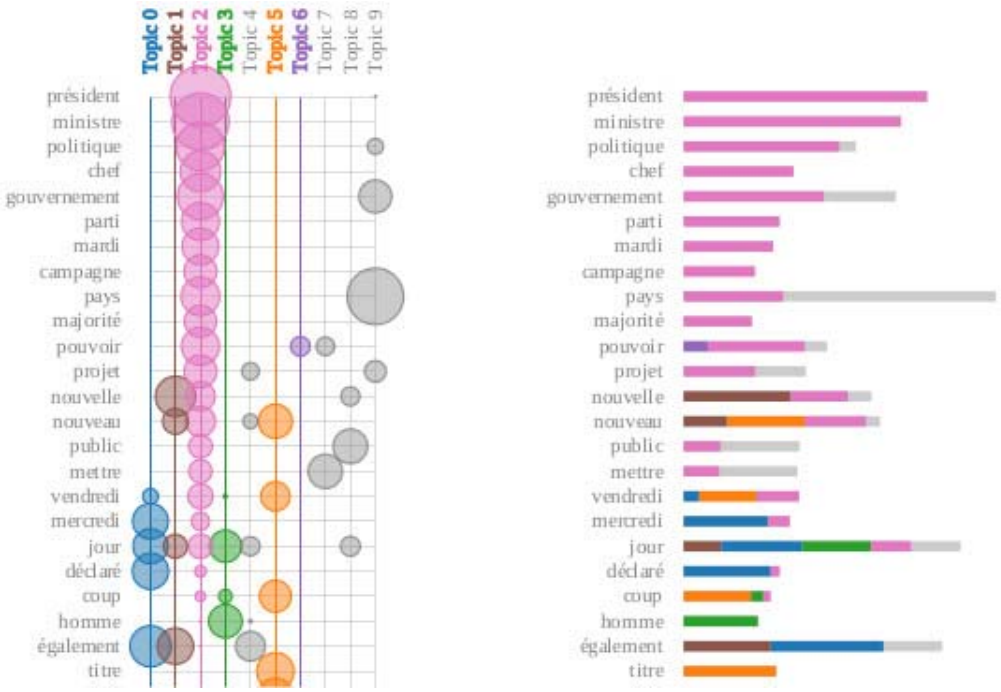

Figure 3 : Visualisation des 10 thématiques obtenues à l'aide de l'outil Termite (Chuang, Manning, & Heer, 2012).

Une fois les thématiques principales déterminées, un texte donné peut être analysé par rapport à ces thématiques. En effet, l'outil Mallet permet également de déterminer dans quelle mesure chacune de ces thématiques est traitée dans le texte individuel sur lequel on choisit de se concentrer. Ce sont ces principes que nous employons actuellement pour l'extraction de traits textuels thématiques, qui sont par la suite associés aux mots inconnus. Plus précisément, chaque mot inconnu est associé à deux textes : d'une part l'ensemble des phrases le comprenant, d'autre part les documents où il apparaît. Les traits que nous allons utiliser sont les proportions des thèmes trouvés par une analyse thématique de ces textes associés aux mots inconnus.

5.3 Résultats obtenus

Les expériences que nous avons effectuées sont basées sur les données décrites au début de cette section : l'objectif est de classer automatiquement les 692 mots/signes inconnus trouvés par notre outil pour déterminer si oui ou non chaque mot représente un candidat néologisme intéressant (avec une grande probabilité). Pour cela nous appliquons une méthode de classification supervisée (SVM, Chang and Lin 2011). Comme une telle méthode est basée sur un apprentissage, redisons-le, chacun des 692 mots inconnus a été examiné par un expert linguiste qui a trouvé que 81 de ces mots étaient de vrais néologismes. À partir de ces données annotées et de traits qui leur sont associés l'outil apprend alors un modèle de classification (un classifieur) qui, par la suite, peut être utilisé pour estimer, pour un mot inconnu, la probabilité que ce mot soit un vrai néologisme.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

Pour nos expériences, nous avons construit des classifieurs à partir des trois groupes de traits présentés dans la première partie de cette section et de leurs combinaisons. Il s'agit de traits basés sur la forme des mots/signes, des traits morpho-lexicaux qui relèvent de caractéristiques morphologiques et lexicales qui, à l'opposé des traits de forme, ne sont pas indépendantes de la langue. Finalement, le troisième groupe de traits que nous étudions concerne les traits thématiques contextuels décrits dans la section précédente.

Pour pouvoir estimer l'impact de ces groupes de traits sur le filtrage des mots/signes nous évaluons nos classifieurs comme suit. Nous effectuons une validation croisée (10-fold cross validation)[14] et présentons également le nombre de vrais néologismes détectés. La Figure 3 montre les résultats de ces expériences.

Au vu de ces résultats la meilleure performance globale est atteinte à l'aide des groupes de traits *forme* et *forme+thème*. Cependant, comme le montre le tableau, le modèle construit dans ces configurations n'a détecté que très peu de vrais néologismes (39 et 43 respectivement). Selon ces résultats, les traits les plus favorables à la détection de néologismes sont les groupes *thème* et *morpho-lex*. On obtient le meilleur équilibre entre une bonne performance globale et un nombre élevé de néologismes identifiés à l'aide du groupe de traits *forme+morpho-lex*.

En utilisant cette procédure, les mots/signes inconnus détectés par notre outil peuvent être reclassés et présentés à l'utilisateur d'une manière plus pertinente, comme le montre l'exemple ci-dessous :

agroécologiste (0.961798)	Etat-départements (0.921507)	...
anti-alcoolisme (0.959942)	nationalistes-révolutionnaires (0.904493)	...
anti-salazariste (0.953199)	démission-surprise (0.891366)	
non-audition (0.939236)	auto-obscurcissant (0.87521)	
multiactivité (0.92963)	constructeur-carrossier (0.870512)	
restaurant-snack-bar (0.925892)	ultra-présent (0.868557)	

Dans cet exemple, on voit que les mots inconnus représentant les candidats néologisme les plus probables (probabilité entre parenthèses) apparaissent automatiquement en tête de la liste, ce qui facilite leur évaluation et leur interprétation. D'autre part, il apparaît que les traits thématiques contextuels ont une incidence fort notable pour la détection automatique de néologismes, en plus de leur importance pour leur observation. En effet, comme il apparaît dans la Figure 3, le groupe de traits *thème* (représentant les thématiques détectées pour le contexte textuel des mots/signes inconnus) permet d'identifier le plus grand nombre de vrais néologismes.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

Classe	Précision	Rappel	F-mesure	Vrais néologismes
pos	0,181	**0,827**	0,297	
pos & nég	0,868	0,548	0,625	**67**

(a) Forme, morpho-lex, thème.

Classe	Précision	Rappel	F-mesure	Vrais néologismes
pos	0,192	0,778	0,308	
pos & nég	0,864	0,597	0,660	63

(b) Forme, morpho-lex

Classe	Précision	Rappel	F-mesure	Vrais néologismes
pos	0,160	0,531	0,346	
pos & nég	0,826	0,625	**0,693**	43

(c) Forme, thème

Classe	Précision	Rappel	F-mesure	Vrais néologismes
pos	0,190	0,481	0,273	
pos & nég	0,832	0,704	**0,751**	39

(d) Forme

Classe	Précision	Rappel	F-mesure	Vrais néologismes
pos	0,132	**0,827**	0,227	
pos & nég	0,836	0,350	0,415	**67**

(e) Morpho-lex

Classe	Précision	Rappel	F-mesure	Vrais néologismes
pos	0,129	**0,889**	0,225	
pos & nég	0,844	0,295	0,338	**72**

(f) Thème

Figure 3 : résultats des expériences pour les classifieurs construits à partir des groupes de traits suivants : forme + morpho-lex + thème, forme+morpho-lex, forme+thème, morpho-lex+thème, forme, thème, morpho-lex. Pour chaque classifieur nous montrons la précision[15], le rappel[16] et la F-mesure[17] obtenus sur la classe positive (les vrais néologismes identifiés correctement) et le résultat global (positif et négatif). Les mesures les plus pertinentes dans le cadre de ces expériences sont le rappel pour la classe positive (la proportion de vrais néologismes détectés) et la F-mesure pour les exemples positifs et négatifs simultanément (performance globale).

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

6 Discussion conclusive

L'objectif de cet article était de montrer que la veille néologique gagne à porter une attention plus grande aux variables textuelles et discursives, ce dont témoignent les études germaniques depuis une trentaine d'années et, sur le plan néographique, ce que font en partie les veilleurs actuels en intégrant des variables contextuelles complémentaires des variables lexicales traditionnelles. En particulier, nous avons mis en relief le fait que les genres de discours, la position textuelle et la thématique sont généralement délaissés alors même que ces variables revêtent des enjeux importants pour l'étude de la néologie en situation journalistique, en particulier, et pour sa compréhension en général.

En l'occurrence, on retiendra qu'une documentation thématique des néologismes autorise un accès à la base qui est complémentaire de l'accès par entrée lexicale, c'est-à-dire par fiche d'attestation. L'enjeu est le suivant : permettre un accès thématique à une ressource néographique c'est ouvrir cette dernière à tous les utilisateurs qui ne sont pas seulement lexicographes ou lexicologues-morphologues. En effet, comme les recherches sociologiques, d'histoire des techniques et les enquêtes journalistiques, par exemple, se penchent toujours sur un sujet particulier, elles ont toutes besoin de constituer leur matière de réflexion autour d'un ou de plusieurs thèmes particuliers. La difficulté du veilleur de néologie consiste ici à se donner une sorte de *répertoire de thèmes* à la fois utile pour la tâche de documentation néographique et pertinent pour les différents types d'utilisateurs visés (sociologue, historien, journaliste, etc.).

De fait, un point important de notre réflexion est que, en situation journalistique, la thématique se laisse concevoir de trois points de vue distincts : 1) comme exploitation dérivée de la variable *domaine lexical*, elle permet d'obtenir un classement les néologismes (cf. la WORTWARTE) ; 2) comme variable textuelle (globale), elle correspond à l'organisation en *rubriques* des journaux (politique, économie, culture, sport, etc.) ; 3) comme variable textuelle (locale), l'analyse thématique permet de caractériser tout texte individuel par une constellation de thèmes principaux et secondaires — telle que la représentent des outils comme ThemeEditor ou Termite. Notre projet de néographie textuelle implique donc la constitution de trois répertoires thématiques distincts.

Le troisième point de vue rencontre un problème qui est spécialement soulevé par l'interprétation des résultats obtenus par les extracteurs de thèmes (ici Mallet). En effet, ces derniers offrent des résultats dont les étiquettes (topics 0, topic 1, etc.), en sortie, sont paradoxalement dénuées d'identité thématique. Comment dès lors qualifier, au moyen d'une dénomination appropriée, un groupe de mots que le calcul statistique considère comme cohérent ? Comment expliciter le thème censé unifier tel ou tel groupe de mots ? Car même l'interprétation d'un résultat du type [Topic 0 = président, ministre, politique, chef, gouvernement, parti], auquel le bon sens attribue l'étiquette *Politique*, requiert de la part du néographe une réflexion prolongée pour parvenir à nommer les traits sémantiques communs aux mots du groupe, au moyen d'un terme générique englobant. Bref, au plan textuel proprement dit, il reste encore à imaginer le pont qui reliera l'extraction automatisée de topics à la documentation thématique des néologismes.

Enfin, l'originalité de notre projet consiste à ne pas réduire le traitement automatisé de la néologie à sa phase initiale de détection : afin de minimiser le temps consacré par le néographe à chaque *fiche d'attestation*, notamment dans le cas particulier d'un veilleur observant journalièrement un corpus de presse, il convient également d'assister la phase finale de documentation des néologismes. Cet objectif pratique de réduction du temps de traitement documentaire constitue un défi technique majeur de la néographie contemporaine. C'est à

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

ce niveau que notre étude produit un résultat très prometteur : l'analyse automatique des thèmes se révèle un atout pour l'identification automatique des néologismes.

Remerciements

Nous remercions Romain Potier-Ferry pour ses contributions à ce travail. Le projet Logoscope est financé par l'Université de Strasbourg dans le cadre de l'Initiative d'Excellence (IdEx) 2012.

Bibliographie

Barz I., Fix U., Schröder M., Schuppener G., eds. (2000). *Sprachgeschichte als Textsortengeschichte. Festschrift zum 65. Geburtstag von Gotthard Lerchner*, Frankfurt/M.

Beust P. (2002). Un outil de coloriage de corpus pour la représentation de thèmes. *In JADT 2002 : 6emes Journées internationales d'Analyse statistique des Données Textuelles*, France, pages 161–172.

Biemann, C., Bordag, S., Heyer, G., Quasthoff, U., & Wolff, C. (2004). Language-Independent Methods for Compiling Monolingual Lexical Data. In A. Gelbukh (éd.), *Computational Linguistics and Intelligent Text Processing*, Vol. 2945,. Berlin, Heidelberg: Springer, pages 217–228.

Blei D.M., Ng A.Y., et Jordan M.I. (2003). Latent dirichlet allocation., *Journal of Machine Learning Research*, 3, pages 993–1022.

Cabré M.-T., Domènech, M., Estopà, R., Freixa, J. Et Solé, E. (2003). L'observatoire de néologie : conception, méthodologie, résultats et nouveaux travaux. In J.-F. Sablayrolles (éd.), *L'innovation Lexicale*, Paris, Honoré Champion, pages 125–147.

Calvet, L.-J. (2010) : *Histoire du français en Afrique*, Paris : Editions Ecriture.

Cartier, E. (2011). Néologie et description linguistique pour le TAL. *Langages*, 183, pages 105–117.

Cartier, E. et Sablayrolles, J.-F. (2008). Néologismes, dictionnaires et informatique. *Cahiers de Lexicologie* 93, pages 175-192.

Chuang, J., Manning, C.D. et Heer, J. (2012). Termite: visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pp. 74-77.

Darmesteter, A. (1888). *La vie des mots, étudiée dans leurs significations*. Paris : Ch. Delagrave.

Dederding, H.-M. (1983). Wortbildung und Text (Zur Textfunktion von Nominalkomposita). *Zeitschrift für germanistische Linguistik* 11, pages 49–64.

Elsen, H. (2004). *Neologismen: Formen und Funktionen neuer Wörter in verschiedenen Varietäten des Deutschen*. Tübingen : Narr.

Elsen, H. et Dzikowicz, E. (2005). Neologismen in der Zeitungssprache. *Deutsch als Fremdsprache*, Leipzig.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

Estopà, R. et Cabré Castellví, M.-T. (2004) *Metodologia del treball en neologia: criteris, materials i processos*, Document de travail, Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada, [En ligne] http://hdl.handle.net/10230/1304 (consulté le 23/11/2013).

Fleischer, W. et Barz, I. (1995). *Wortbildung der deutschen Gegenwartssprache*. Tübingen : Max Niemeyer.

Gérard, C. (2011). Création lexicale, sens et textualité. *PhiN, Philologie im Netz, 55-2*.

Gérard, C. et Kabatek, J. (2012). La néologie sémantique en question : quelles conceptions pour quelles méthodes ? In Kabatek, J. et Gérard, C. (éd.), *Néologie sémantique et corpus : méthodes statistiques*, numéro thématique des *Cahiers de lexicologie*, Paris : Editions Garnier.

Glessgen, M.-D. (2012). *Linguistique romane*. Armand Colin.

Hearst, M.A. (1997). TextTiling: segmenting text into multi-paragraph subtopic passages, *Computational Linguistics*, 23, 1, pages 33–64.

Kastovsky, D. (2006). Morphology as word-formation in the 20th-century linguistics: A survey. In Auroux, S., Koerner, K., Niederehe, H.-J. et Versteegh, K. (éd.), *Manuel international d'histoire des études linguistiques des origines à nos jours*, Vol. 3, pages 2324–2340.

Kerleroux, F. (2006). Les théories morphologiques à la fin du XX^{e} siècle. In Auroux, S., Koerner, K., Niederehe, H.-J. et Versteegh, K. (éd.), *Manuel international d'histoire des études linguistiques des origines à nos jours*, Vol. 3, pages 2313–2324.

Koch, P. et Oesterreicher, W. (1985) Sprache der Nähe - Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch 36/85*, pages 15-43.

Koch, P. (1997). Diskurstraditionen: zu ihrem sprachtheoretischen Status und ihrer Dynamik. In Frank, B. et alii (éd.), *Gattungen mittelalterlicher Schriftlichkeit*, Tübingen, pages 43–79.

Koch, P. et Oesterreicher, W. (1990). *Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanisch* (= Romanistische Arbeitshefte 31). Tübingen : Niemeyer, pages 5-17.

Lemnitzer, L. (2010) : Neologismenlexikographie und das Internet. *Lexicographica, 26*, pages 65–78.

Loiseau, S. (2012).Un observable pour décrire les changements sémantiques dans les traditions discursives : la tactique sémantique », *Cahiers de lexicologie*., 2012-1, 100, pages 185–199.

Matussek M. (1994) : *Wortneubildung im Text, Beiträge zur germanistischen Sprachwissenschaft Bd. 7*, Hamburg : Helmut Buske.

Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol, I. et Nouvel, D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées . *Traitement Automatique des Langues*, 52 (1), pages 69-96.

Oesterreicher, W. (1997). Zur Fundierung von Diskurstraditionen. In Frank, B., Haye, T., Tophinke, D. (éd.), *Gattungen mittelalterlicher Schriftlichkeit*, Tübingen : Narr (ScriptOralia, 99), pages 19-41.

Mccallum, A., Mimno D.M. et Wallach, H.M. (2009). Rethinking LDA: Why Priors Matter. In *Advances in Neural Information Processing Systems*, pp. 1973-1981.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

Medelyan, O. et Witten, I.H. (2008). Domain-Independent Automatic Keyphrase Indexing with Small Training Sets., *Journal of the American Society for Information Science and Technology*, 59, 7, pages 1026–1040.

Ollinger, S. et Valette, M. (2010). La créativité lexicale : des pratiques sociales aux textes. *Actes del I Congrés Internacional de Neologia de les llengües romaniques (CINEO'08)*, M. Teresa Cabré i Castellví et alii (éd.), Publicacions de l'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF), pages 965–876.

Peschel, C. (2002). *Zum Zusammenhang zwischen Wortneubildung und Textkonstitution*. Tübingen : Niemeyer.

Quasthoff, U., Richter, M. et Biemann C. (2006). Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the fifth international conference on Language Resources and Evaluation, LREC*, pp. 1799-1802.

Romary, L., Salmon-Alt, S., et Francopoulo, G. (2004). Standards going concrete: from LMF to Morphalou. *Workshop On Enhancing And Using Electronic Dictionaries at the 20th International Conference on Computational Linguistics - COLING 2004*.

Sablayrolles, J.-F. et Pruvost, J. (2003). *Les néologismes*. Paris : PUF.

Sablayrolles, J.-F. (1996–1997). Néologismes : une typologie des typologies . *Cahiers du C.I.E.L.,* U.F.R. E.I.L.A., Paris-7, pages 11–48.

Sablayrolles, J.-F. (2011). Neologia : un dictionnaire néologique sous forme de base de données. In Marcelino Cardoso, S. A., Mejri, S. et Mota, J. A. (éd.), *Os di.ci.o.na.rios, fontes, métodos et novas tecnologias*, Instituto de Letras da Universidade federal da Bahia, Brésil, pages 221-235.

Schlieben-Lange, B. (1983). *Traditionen des Sprechens. Elemente einer pragmatischen Sprachgeschichtsschreibung,* Stuttgart : Kohl- hammer.

Siebold, O. (2000). *Wort - Genre - Text. Wortneu-bildungen in der Science Fiction.* Tübingen : Gunter Narr Verlag.

Siebold, O. (2005). Die Sprache der Science Fiction. *Der Deutschunterricht* 2, pages 69-73.

Steyvers, M. et Griffiths, T. (2007) *Probabilistic Topic Models*. Lawrence Erlbaum Associates.

Wilhelm, R. (2001). Diskurstraditionen. In Haspelmath, M., König, E., Oesterreicher, W. et Raible, W. (éd.), *Language Typology and Language Universals. An International Handbook*, I, Berlin/New York, de Gruyter, pages 467–477.

Wilhelm, R. (2003). Von der Geschichte der Sprachen zur Geschichte der Diskurstraditionen. Für eine linguistisch fundierte Kommunikationsgeschichte. In Aschenberg, H. Wilhelm, R. (éd.), *Romanische Sprachgeschichte und Diskurstradition. Tübingen*, pages 221–236.

[1] Les veilleurs de néologie non seulement comportent une phase de détection des néolgismes (semi-automatisée), identifiés au sein d'un corpus dynamique méthodologiquement défini (par ex. un ensemble de six journaux français en ligne) et une phase de documentation qui approvisionne la base de néologismes proprement dite, qui est une sorte de ressource lexicale. Les veilleurs se distinguent donc à la fois des bases constituées intuitivement « à l'œil et à la main »

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

(ex. Borneo, http://web.atilf.fr/BORNEO-Base-d-Observation-et-de) et des purs outils d'extraction qui n'ont eux pas intégrés comme la composante d'une entreprise néographique plus ambitieuse. Parmi ces outils, on peut citer l'extracteur Pompamo (Ollinger et Valette 2010) qui, à partir du corpus étiqueté qu'on lui soumet, note en sortie l'orthographe de chaque néologisme candidat, sa nature morphosyntaxique (nom, verbe, etc.) ainsi que son genre grammatical et son nombre.

[2] La fiche, pour chaque entrée, se borne à indiquer la marque du génétif, celle du pluriel et le domaine lexical (éducation, santé, télécommunication, environnement, etc.) (http://www.wortwarte.de).

[3] neologia propose en outre une définition de l'entrée décrite et endosse ainsi la tâche centrale du lexicographe. Cette base se donne aussi la possibilité de commenter l'origine historique du mot, en signalant par exemple que la création est due à une succession de faits divers (ex. [chien] mordeur). C'est là une différence de taille avec notre projet XXX qui, concernant la pratique documentaire, se borne lui à s'adresser au lexicographe et à l'historien de la langue en leur fournissant une information aussi brute et peu interprétée que possible. Notre choix réduit bien entendu le temps de description consacré à chaque néologisme et rend ainsi plus réaliste un enrichissement journalier de la ressource.

[4]La variable domaine lexical est une constante des dictionnaires et des veilleurs de néologie. Le Petit Robert possède son propre répertoire de domaines (admin., inform., lang., telecomm., etc.). Les veilleurs wortwarte et obneo documentent eux, respectivement, les catégories « thématique » (on y reviendra) et « relatif à » (une aire géographique, un nom de personne, une organisation politique, une entité sportive, etc.).

[5]Nommés « matrices » les types de formation retenus sont formels (sufixation, composition, etc. ; Sablayrolles, 1996 et1997, Sablayrolles et Pruvost, 2003) et sémantique (métaphore / métonymie / autres). Actuellement toutefois, pour une raison technique d'interface, neologia ne permet pas de documenter l'usage simultaé de plusieurs matrices (par ex. escorteuse est à la fois une suffixation et un euphémisme). Quant à la variable influence linguistique, elle permet de consigner la langue concernée (anglais, arabe, etc.) ainsi que le mode d'influence (traduction, calque, emprunts, création d'équivalent).

[6]Comme n'importe quelle unité de langue, toute création lexicale se définit par une identité variationnelle, marquée ou non marquée : variation diatopique (situation géographique : français méridional, français du Sénégal), variation diastratique (« niveaux de langue » : français spécialisé / cultivé / moyen / populaire), variation diaphasique (« styles de langue » : français soutenu / usuel / familier / vulgaire).

[7]Dirigé par P. Blumenthal (Université de Cologne) et S. Mejri (Université Paris 13), http://syrah.uni-koeln.de/varitext/.

[8]On n'étudie dès lors plus séparemment le lexique et le texte : comme les genres de discours déterminent l'usage de la variation linguistique (Koch et Oesterreicher, 1985 et 1990 ; Glessgen, 2012 : 124-127), la prise en compte des genres comme variable néographique permet d'indiquer des tendances du texte à favoriser/interdire l'usage d'une diaphasie haute/basse, par exemple.

[9]OBNEO indique le domaine de compétence ou de savoir de la source dans son corpus d'écrits spontanés (i.e. documents écrits autres que la presse). Ainsi, étonnement, par « type de texte » (cat. tipus de text) OBNEO n'entend pas une sorte de texte particulier (texte de loi, tract de propagande, revue universitaire, etc.), mais un ensemble de publications associé à un certain domaine professionnel (économie, sport, culture, informatique) ou sans domaine privilégié (publications d'intérêt général). Par exemple Dir Fitness relève du type « sport », Teatre BCN o Marges du type « culture », etc. Dans sa documentation du contexte, Neologia comporte une variable « domaine » de la source aux étiquettes analogues : économie, sport, société, politique, culture, etc. Ce domaine de compétence de la source ne doit pas être confondu avec le domaine lexical du néologisme.

[10]Obneo oberve également un corpus oral issu d'émissions de radio et de télévision (Catalunya Ràdio, COM Ràdio, Televisió de Catalunya, Ràdio 9). Outre les variables attendues pour le type de textes oraux (transcription phonétique, minute d'apparition, etc.), pour s'adapter à l'oralité de ce corpus, OBNEO documente des caractéristiques socio-linguistiques : rôle de l'émetteur (présentateur, reporter, etc.), âge et sexe de l'émetteur, dialecte (català oriental, català balear, etc.) et langue maternelle (castillan, anglais, français, etc.).

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801208

[11]La littérature sur le sujet est devenue importante en romanistique, et pas seulement en quantité : e.g. Schlieben-Lange, 1983 ; Koch, 1997 ; Oesterreicher, 1997 ; Wilhelm, 2001et 2003.

[12]Le style collectif singularise parfois nettement un journal : par exemple, les jeux de mots fréquemment utilisés dans les titres de Libération sont un indice de style rédactionnel.

[13]Term Frequency Inverse Document Frequency est une méthode de pondération des termes fréquemment utilisée en recherche d'information.

[14]Pour cela l'ensemble de 692 mots/signes est divisé aléatoirement en 10 échantillons, l'apprentissage s'effectue sur 9 de ces 10 échantillons et la performance du modèle est mesuré sur l'échantillon restant. Le résultat final de l'évaluation est la moyenne des performances des 10 modèles qu'on peut obtenir de cette façon.

[15]La précision est la proportion des néologismes correctement identifiés par rapport au nombre total de néologismes détectés par notre outil.

[16]Le rappel est la proportion des néologismes correctement identifiés par rapport au nombre total de néologismes validés par l'expert linguiste.

[17]La F-mesure combine la précision et le rappel : F = 2*(rappel * précision)/(rappel + précision)