NeoTrack – Un analyseur de néologismes en ligne

Maarten Janssen
IULA / ILTEC
maarten.janssen@upf.edu

Abstrait

NeoTrack est un outil pour la détection semi-automatique de néologismes dans les textes électroniques. Il produit automatiquement une liste de candidats à néologismes et fournit une interface fonctionnelle pour le traitement manuel de ces candidats. NeoTrack est un outil en ligne, avec lequelle les chercheurs peuvent collaborer depuis différentes villes, différents pays et continents, tout en disposant des mêmes données. NeoTrack est un programme de sources ouverts, s'adapte à la plupart des serveurs et est facile d'installer. Pour utiliser NeoTrack, la seule chose indispensable est une liste des mots de la langue avec laquelle on désire travailler. Les outils complexes de traitement automatique du langage ne sont pas nécessaires.

1. Introduction

L'étude des mots nouveaux (néologismes) dans une langue est une activité utile, réalisée à de nombreuses fins. Traditionnellement, la détection et compilation de néologismes fait partie du domaine de la lexicographie, en cherchant des mots nouveaux afin d'actualiser la nomenclature des dictionnaires. Ce travail, manuel et laborieux, consistait en l'examen d'un grand volume de textes pour le but de relever des mots ou expressions jugés nouveaux, bien parce qu'ils n'étaient pas utilisés auparavant (dans le cas de néologismes formels), bien parce qu'ils sont utilisés avec un sens nouveau (dans le cas des néologismes sémantiques).

Au cours de la dernière décennie, la détection de néologismes a beaucoup évoluée. Elle n'appartient plus au domaine exclusif du lexicographe; actuellement, de nombreux centres de recherches linguistiques étudient la néologie dans d'autres buts, comme, par exemple, l'étude des processus morphologiques productifs d'une langue. De plus, une grande partie de la détection de néologismes est maintenant fait (semi-) automatiquement.

Actuellement, il existe un grand nombre de programmes pour le traitement automatique de néologismes, tels que AVRIL et AVIATEUR (Renouf, 1993), SEXTAN (Vivaldi, 2000), ZeitGeist (Veale, 2006), Telenaute (Isaac, 2008), etc. Ces applications peuvent être divisées dans deux catégories : les systèmes entièrement automatiques et les semi-automatiques. Les systèmes entièrement automatiques emploient une combinaison de consultations lexicales et d'analyses statistiques (ou seulement d'analyses statistiques) afin d'extraire une liste de néologismes d'un texte, sans l'intervention d'un utilisateur humain. Ces systèmes ont pour but une rappel élevé (produire tous les néologismes) et une précision élevé (produire seulement des néologismes), étant l'inévitable omission d'une certaine quantité de néologismes par le système et des mots, dans las liste de résultats, n'étant pas des néologismes.

D'autre part, les systèmes semi-automatiques extraient une liste de néologismes potentiels (ou de *candidats à néologismes*) d'un texte et dépendent de l'intervention humaine pour décider si les candidats sont, ou ne sont pas, des néologismes. NeoTrack est un système répondant à ces caractéristiques, créé pour l'*Observatório de Neologia de Português* (ONP), projet de ILTEC en Portugal (Correia *et al.*, 2005). NeoTrack a été conçu comme un outil léger nécessitant peu de ressources externes et, par conséquent, peu de ressources spécifiques pour une langue donnée. Il est accessible librement à travers l'Internet. Cela fait de lui un outil facile à

implémenter dans n'importe quel langue et ceci sans la nécessité d'utiliser des outils complexes pour le traitement automatique du langage.

Cet article présente le système NeoTrack à travers trois chapitres: dans le chapitre 2, le système est décrit depuis la perspective de l'utilisateur. Le chapitre 3 présente le système depuis une perspective plus technique et explique son fonctionnement interne et les réquisits pour son utilisation. En conclusion, le chapitre 4 offre un contexte théorétique, une comparaison avec quelques systèmes de caractéristiques semblables, de même que des développements futures possibles (et impossibles).

2. Interface

NeoTrack est un système entièrement en ligne, dans lequel toute l'interaction entre l'utilisateur et le serveur se réalise à travers un navigateur de l'Internet. Il est possible d'accéder à NeoTrack de n'importe où, à partir du moment où l'on dispose d'un ordinateur connecté à l'Internet. Cette caractéristique permet la collaboration de différentes universités sur le même corpus. Neotrack permet également aux chercheurs de travailler depuis leur domicile ou en déplacement. L'accès au système est limité par un identifiant/mot de passe.

NeoTrack détecte et procède des néologismes (potentiels) dans des textes en format électroniques. La méthode à partir de laquelle cette détection est réalisée est la suivante : d'abord, l'utilisateur télécharge un fichier dont il veut extraire les néologismes. Puis, le système produit une liste de tous les mots dans ce fichier qu'elle ne connaît pas, ou bien la liste de candidats à néologismes. Après de cela, les candidats à néologismes doivent être classifiés manuellement comme de véritables néologismes ou bien des candidats erronés. Les candidats jugés néologiques sont stockés dans une base de données qui peut être consultée et modifiée en tout moment. Ce chapitre décrit ces différentes parties de NeoTrack en séquence, aussi que l'accès (facultatif) pour les internautes à la base de données.

2.1 Gestion des fichiers

Étant donné que NeoTrack fonctionne sur un serveur, la première étape consiste à télécharger les fichiers que l'on veut traiter. Les fichiers téléchargés sont gardés dans un dossier temporaire, en attendant leur traitement. Ces fichiers temporaires peuvent être consultés afin de vérifier s'il y a eu des erreurs au cours du transfert. Le système permet également d'extraire des informations relatives à ces fichiers, comme le nombre total de mots ou la fréquence d'apparition des mots dans le fichier (voir figure 1).

Une fois que les fichiers sont téléchargés, il est possible de les traiter. Le traitement d'un fichier signifie que le système nettoie le texte, extrait la liste des mots du texte et produit une liste des mots inconnus - les candidats à néologismes. Ce processus inclut des filtres pour éliminer les séquences de lettres qui ne sont pas des mots (noms propres, adresses courriel, etc.)

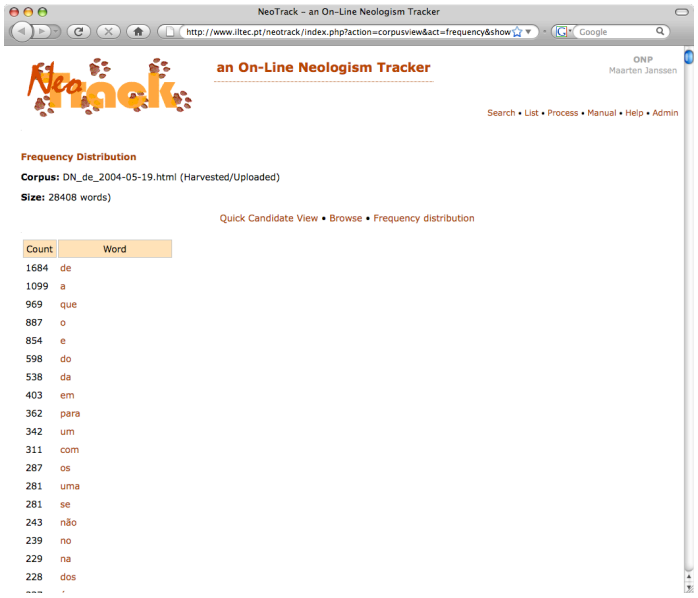

Figure 1. Page avec la répartition de fréquences

La liste des candidats à néologismes résulte de la soustraction de la liste de l'ensemble des mots du texte de tous les mots connus de la langue. Pour réaliser cette opération, NeoTrack utilise une liste simple contenant tous les mots dont on sait que ce sont des mots de la langue. Cette liste s'appelle la *liste d'exclusion*.

Une fois la liste des candidats générée, les autres mots du texte et les mots de la liste d'exclusion ne font plus partie du processus. Lorsque l'on utilise une liste d'exclusion qui peut varier (voir paragraphe 3.3), il est important de l'actualiser. Sinon, les mots sont ajoutés a la liste d'exclusion seraient inclus dans la liste des candidats à néologismes.

Tous les textes traités avec NeoTrack sont gardés dans le serveur (sauf s'ils sont pas effacés intentionnellement). NeoTrack utilise la version nettoyée des textes pour tous son fonctionnement interne, mais parce que la typographie des mots est souvent importante pour la recherche de néologisme, les versions originales des textes sont maintenues sur le serveur aussi. La totalité des textes constitue un corpus qui croit graduellement. Il est possible d'exploiter ce corpus en utilisant, par exemple, le concordancier SimpleConc (Janssen & Freitas, 2008). Il est prévu que SimpleConc soit prochainement intégré dans NeoTrack afin fournir cette fonction par défaut.

2.2 Classification des candidats

Le module principal de NeoTrack est la classification des candidats à néologismes. NeoTrack contient une base de données de tous les candidats à néologismes rencontrées dans tous les fichiers traités par le système. Il est possible de consulter la liste complète des candidats pour un fichier donné : le système indique le statut du candidat (ouvert/complet) de même que la classification de chacun des candidats. Ce type de consultations permet, entre autres, d'obtenir le nombre total de candidats traités, la distribution des modes de traitement et d'inclure comme néologisme un candidat accidentellement rejeté.

NeoTrack fournit une page web pour chaque candidat à néologisme contenant toute l'information nécessaire pour sa classification. Un exemple d'une tel page est donné dans la figure 2. Les données de cet exemple sont extraites du projet ONP.

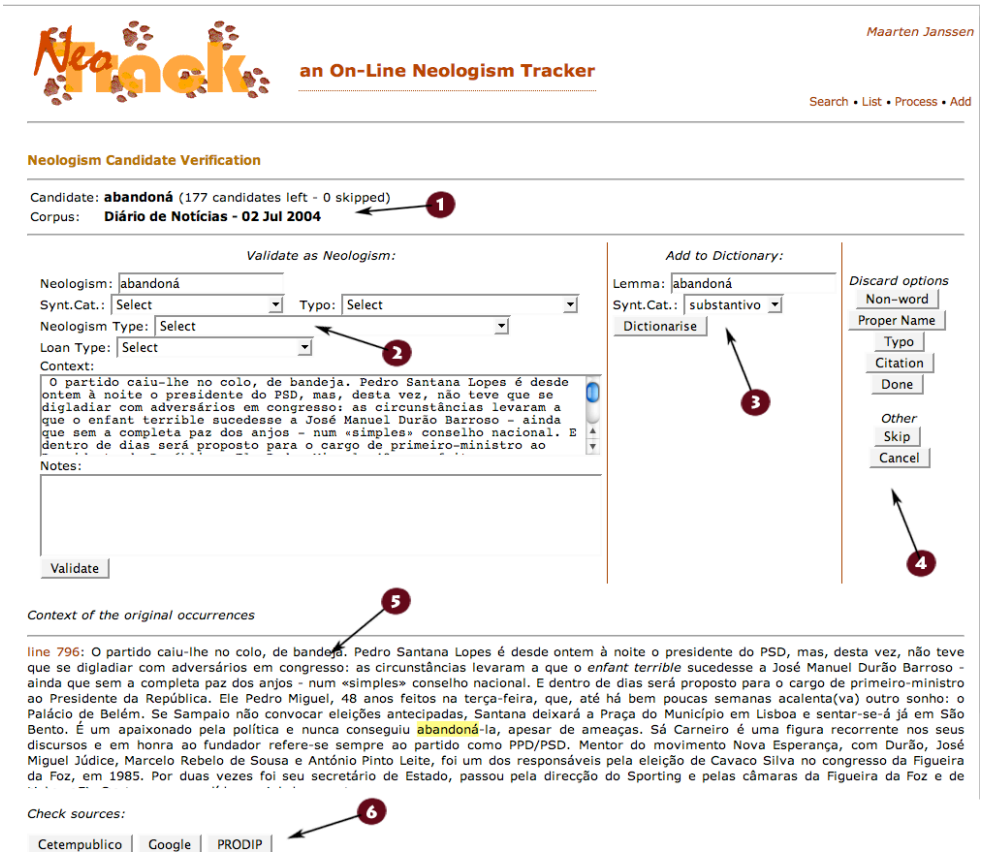

Figure 2. Classification des candidats à néologismes

La partie supérieure de la page (correspondant au numéro 1 dans l'illustration) donne d'information sur le candidat: la graphie dans laquelle elle se trouve dans le texte (*abandoná*), le nom et la date du source dans lequel on l'a trouvé (le journal *Diário de Notícias* du 2ème du juillet 2004), et la quantité de candidats à néologismes dans ce source qui n'ont pas été encore classifiés (177).

La partie inférieure de la page (5) montre le(s) contexte(s) du mot dans le fichier original. Pour faciliter la lecture, le candidat est souligné dans le texte. Pour les candidats qui apparaissent plusieurs fois dans le même texte, tous les contextes du mot sont listés. La taille du contexte dépend du fichier original (longueur d'une ligne). Il est possible que le contexte soit insuffisant pour juger le candidat. Dans ce cas, il est possible de cliquer sur le numéro montré au début du contexte. Le fichier complet s'ouvre en indiquant la ligne appropriée.

Quand le candidat est jugé néologique, on peut l'ajouter à la base de données de néologismes grâce au formulaire de validation (2). Ce formulaire, demande à l'utilisateur de fournir diverses informations sur le néologisme, comme la forme de citation, la catégorie syntactique, le type de néologisme, etc. (voir section 2.3). Dans la mesure du possible, l'information est remplie automatiquement par le système, mais toujours en offrant toujours la possibilité à l'utilisateur d'éditer. Le système choisit par défaut la première occurrence du mot comme contexte du néologisme. Mais le contexte peut être augmenté, raccourci ou changé par un autre contexte si le candidat apparaît plusieurs fois dans le texte.

Il y a deux raisons pour rejeter un candidat comme néologisme. La première raison pourrait être due au fait qu'il ne s'agisse pas d'un mot : erreur orthographique, nom propre, partie d'une citation dans une langue étrangère, adresse email. Grâce aux boutons du formulaire d'écart (4), il est possible de supprimer de tels mots en un seul clic, ainsi que d'indiquer la

raison de leur rejet. Il est aussi possible de remettre la classification du néologisme (*skip*), par exemple quand il est difficile de classifier le candidat, et précéder au prochain candidat.

La deuxième raison pour rejeter un néologisme pourrait être due au fait qu'il s'agisse d'un mot qui fait partie du lexique de la langue mais, pour diverses raisons n'était pas inclus dans la liste d'exclusion. Pour vérifier s'il s'agit bien d'un tel trou dans la liste d'exclusion, l'option en bas de la page (7) peut fournir diverses liens aux corpus en ligne, dans lesquels le candidat peut être recherché. Dans le projet ONP, le corpus *CETEMPublico* et le corpus de *Centro de Linguística de Universidade de Lisboa* (CLUL) sont utilisés comme corpus de référence. Quand le mot a été vérifié dans les corpus de référence avec une fréquence suffisamment élevée, on peut l'ajouter à un dictionnaire provisoire, en indiquant sa catégorie syntactique de même que d'autres informations. Les mots du dictionnaire provisoire peuvent être ajoutés à la liste d'exclusion directement, ou au moyen d'une base de données plus raffinée nommé OSLIN (voir la section 3.3).

2.3 Gestion de la base de données

Les néologismes accumulés avec NeoTrack et l'information qui leur est relative sont stockés dans une base de données. Les différents champs de cette base de données de néologismes proviennent du projet NEOROM (Cabré, 2006) : le *néologisme* (dans sa forme de citation), la *catégorie grammaticale*, la *typographie* dans laquelle il apparaît (gras, italique, guillemet, etc.), la *type de néologisme* (sémantique, pragmatique, emprunt, dérivation, etc.), la *type d'emprunt* (pour les emprunts, la langue d'origine), *contexte*, *source du contexte*, *date*, et *éditeur* (l'utilisateur qui a entré le néologisme). Il n'est pas nécessaire d'utiliser tous ces champs, et la liste de valeurs possibles pour chacun d'eux peut être définie pour chaque langue.

Il est possible d'exporter tous les néologismes compilés avec NeoTrack vers un fichier, et de les utiliser dans le logiciel de préférence. Mais il est aussi possible de réaliser la gestion complète des néologismes avec NeoTrack. En utilisant le système de gestion interne, il est possible d'examiner ou de chercher l'ensemble des néologismes et de corriger les possibles erreurs.

Dans la base de données, aux de néologismes compilés semi automatiquement, il est possible d'ajouter des néologismes détectés manuellement. Ainsi, NeoTrack peut être utilisé comme banque de donnés unique contenant des néologismes de types différents. Pour l'addition manuelle de néologisme, NeoTrack fournit un formulaire analogue au formulaire (2) de l'illustration 2, dans laquelle toute l'information relevante peut-être ajouter manuellement (forme de citation, catégorie grammaticale, etc.). Ainsi, des néologismes détectés dans des formats non-électroniques, comme le papier, de même que des néologismes sémantiques et des expressions multi-mots néologistiques ne pouvant être capturées avec NeoTrack (voir section 4.1) peuvent être ajoutés. Étant donnée la différence de statut entre les néologismes traités à la main et les néologismes traités semi automatiquement, toutes les entrées de la base de données ont une marque indiquant leur mode de détection.

2.4 Accès pour télénautes

Le système NeoTrack étant entièrement basé sur l'Internet, il est facile de fournir accès à la base de données de néologismes a tous les utilisateurs de l'Internet. Les télénautes peuvent examiner et chercher les néologismes, que ce soit sur une liste ou dans une page contenant toute l'information détaillée relative à un néologisme. L'interface de l'accès pour télénautes peut être personnalisée. Un exemple de la page du projet ONP est donné dans la figure 3.

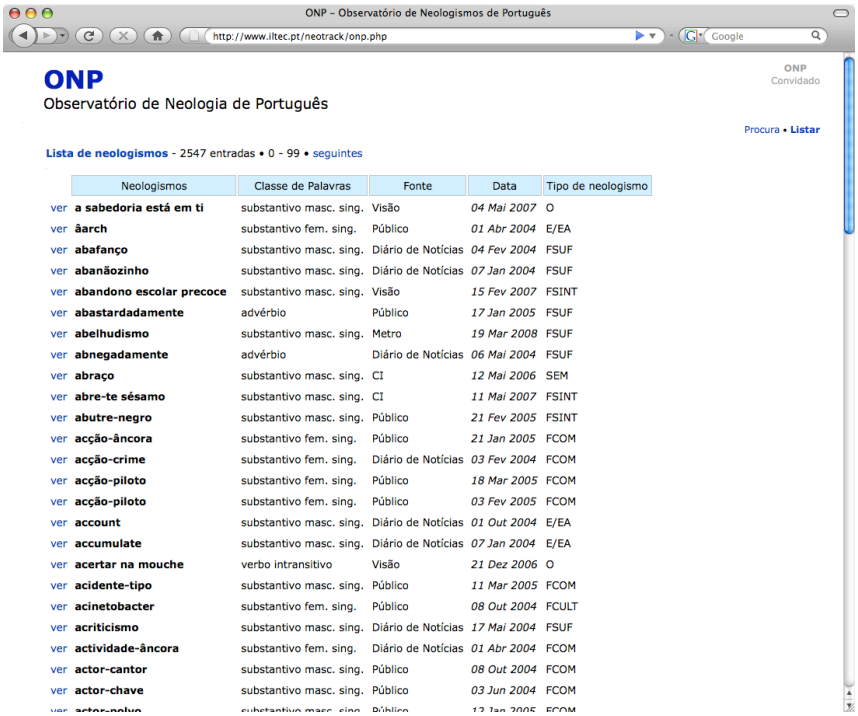

Figure 3. Accès pour télénautes du projet ONP

En utilisant NeoTrack, il est possible de donner accès à tous les télénautes, de désactiver complètement l'accès des télénautes, ou encore de donner accès à un nombre spécifique d'utilisateurs à travers l'utilisation d'un mot de passe. La gestion des différents utilisateurs est réalisée par un système de permissions données par un *administrateur*. L'administrateur peut créer de nouveaux utilisateurs, changer les mots de passe des utilisateurs existants, ou changer leur statut. Il existe trois types de statuts : les utilisateurs *télénautes*, qui peuvent seulement voir les néologismes ; les utilisateurs *éditeurs*, qui peuvent voir et éditer les néologismes, et traiter des textes digitaux ; et les utilisateurs *désactivés*, qui sont inscrits dans le système mais n'y ont plus accès.

Comme il est possible qu'il y ait des inconsistances et des erreurs dans les néologismes récemment compilées, toutes les entrées de le base de données de néologismes ont une marque de « vérification », qui permet d'indiquer si le néologisme a été vérifié ou pas. L'accès pour télénautes permet de visualiser seulement les néologismes vérifiés.

3. Description technique

NeoTrack est un logiciel léger, de sources ouvertes et de facile application pour toutes les langues, étant donné qu'il ne dépend pas d'outils externes pour le traitement des néologismes. Le système utilise un modèle serveur/client, dans lequel le logiciel est situé sur le serveur et les clients se branchent sur le serveur avec un navigateur Internet. Ce chapitre décrit la configuration requise pour utiliser NeoTrack et donne une vision plus technique de la méthode avec laquelle la liste de candidats à néologismes est créée.

3.1 Configuration requise

NeoTrack se base sur un modèle serveur/client, il y a donc des réquisits pour les clients et des réquisits pour le serveur, bien qu'ils soient tous les deux sont assez limités. En ce qui concerne les clients, le système utilise seulement du HTML simple ce qui équivaut à dire qu'il est possible d'utiliser NeoTrack depuis n'importe quel ordinateur connecté à Internet. Il

peut être utilisé avec n'importe quel navigateur bien que le système ait été testé avec FireFox et Internet Explorer.

En ce qui concerne le serveur, le système utilise uniquement des outils UNIX standards: PHP, MySQL, Perl et des commandes basiques. Il doit être possible d'utiliser NeoTrack avec n'importe quel serveur LINUX ou UNIX, sans la nécessiter d'installations additionnelles. Pour utiliser des fonctions avancés, comme l'usage des fils RSS (voir 3.2), diverses permissions du serveur sont nécessaires, par exemple pour la création des commandes périodiques (cron commands). En dehors du processus de création de la liste de néologismes, NeoTrack n'utilise pas beaucoup de temps du processeur, et le système devrait même fonctionner dans les serveurs les plus anciens. Bien que tous les fichiers traités soient gardés dans le serveur, il est nécessaire d'avoir un minimum d'espace disponible dans le disque dur.

3.2 Fonctionnement interne

Le schéma du fonctionnement interne de NeoTrack est illustré dans la figure 4. Le système extrait les candidats à néologismes d'un texte brut (*corpus text*) en deux étapes. La première étape consiste en la création d'une liste de tous les mots qui apparaissent dans le texte (*corpus words*). La seconde étape est une comparaison de cette liste avec la liste de tous les mots connus de la langue (*exclusion list*). La liste résultante des mots (*neologism candidates*) est présentée au linguiste pour différencier les unités jugées néologiques des autres.

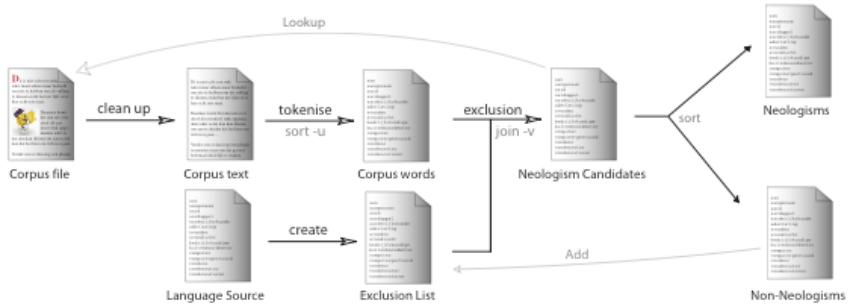

Illustration 4. Schéma Technique de NeoTrack

La liste des mots du corpus est créée par une script de tokenisation en Perl. L'extraction de la liste d'exclusion est réalisée par le logiciel en ligne de commande *join*. Et les pages pour la classification des candidats (décrite au paragraphe 2.2) sont gérées par plusieurs scripts en PHP qui produisent des pages de classification (en HTML simple).

NeoTrack effectue l'extraction de néologismes de textes électroniques, provenant d'Internet. NeoTrack permet donc d'utiliser des textes en HTML (*corpus file*), tirant le texte simple d'eux. Le système ne permet pas l'extraction de texte d'autres types de fichier, comme Word ou PDF parce que, bien qu'une telle extraction soit possible, ce type de extraction n'est pas très fiable. Pour cette raison, les textes doivent être en HTML ou texte brut, en codage ISO-8859-1 (les textes HTML sont automatiquement recodifiés).

La liste d'exclusion de NeoTrack est gardée comme une simple liste de mots. Le système fonctionne avec n'importe quelle liste de mots, quelle que soit son origine. Mais l'idée de fond c'est qu'on utilise un dictionnaire comme source pour la liste d'exclusion initiale pour faire fonctionner NeoTrack. Bien qu'une liste simple des mots suffise, il est néanmoins possible d'utiliser une base de données lexicale plus structurée afin d'en extraire la liste d'exclusion. La base de donnée lexicale interne de NeoTrack (décrite dans la section suivante) est nommée OSLIN. En utilisant OSLIN, le système produit automatiquement une nouvelle liste d'exclusion chaque fois que l'on est en train de traiter un nouveau texte, contenant tous les mots de la base de données.

La dernière version de NeoTrack permet d'utiliser des fils RSS pour collecter des journaux au lieu de télécharger des textes individuels. Plusieurs journaux n'ont plus une version HTML de leur journal, mais fonctionnent avec des dossiers PDF en ligne. Cependant, la plupart d'entre eux fournissent leur contenu dans des fils RSS. Utilisant le système RSS, NeoTrack utilise une base de données qui contient tous les articles apparaissant sur le fil RSS. Avant la détection de néologismes, tous les articles correspondant à une même journée sont alors rassemblés dans un fichier HTML. Ce fichier est ensuite traité comme fichier téléchargé.

3.3 OSLIN

La base de données lexicale OSLIN (Open Source Lexical Information Network) est un système modulaire, basée sur une structure relationnelle. Le cœur du système est constitué de deux tables, l'une composée d'unités lexicales, l'autre de formes flexionnelles (Janssen, 2005). Les unités ont un identificateur unique, une forme de citation, une catégorie grammaticale, ainsi que diverses données optionnelles : structure syllabique, genre, etc. Les formes flexionnelles ont une orthographie, un code pour indiquant quelle forme est que c'est (singulier, passé composé, etc.), et un lien renvoyant à l'entrée correspondante.

Ces deux tables forment le cœur d'un réseau de ressources lexicales. Tous ces ressources sont des tables individuelles, liées aux entrées lexicales, aux formes flexionnelles, ou bien à une combinaison de tables variée. Pour le Portugais, la base de données *MorDebe*, développé à l'ILTEC à Lisbonne, contient entre autres des tables pour l'information grammaticale des mots (types de verbe, types de nom, etc.), pour des relations dérivationnels (entre les verbes et leurs noms dérivationnels, entre les noms et leurs superlatifs, etc.), pour la transcription phonétique, pour les toponymes, les gentilés, et pour la structure morphologique. Le contenu de MorDebe est consultable sur Internet, sur le *Portal da Língua Portuguesa* (http://www.portaldalinguaportuguesa.org).

OSLIN dispose de son propre système de gestion, basé lui aussi sur Internet (Barbosa *et al.* 2008). Avec le système de gestion, il est possible d'ajouter, d'éditer, ou d'effacer des entrées lexicales, avec tous les formes flexionnelles qui lui sont associées. Quand il y a un logiciel intégré dans OSLIN pour la création de paradigmes (comme c'est le cas pour le portugais), le système produit automatiquement toutes les formes flexionnelles chaque fois qu'on ajoute une nouvelle entrée lexicale. Les formes produites sont préremplies dans le formulaire pour ajouter un mot et sont toujours vérifiées par l'utilisateur. Étant donné que les paradigmes produits ne sont jamais utilisés sans vérification, il n'est pas nécessaire que le logiciel qui les produit soit exempt d'erreurs, ce qui rend la création d'un tel logiciel beaucoup plus facilement réalisable.

Il est possible d'utiliser OSLIN en combinaison avec NeoTrack. Dans ce cas, les deux systèmes sont intégrés dans les deux directions: d'un côté, OSLIN fonctionne comme le stockage plus structuré des mots de la liste d'exclusion, en représentant les formes liées à leur entrée lexicale. Juste après de traiter un nouveau texte, le système produit optionnellement une nouvelle liste d'exclusion, contenant toutes les formes se trouvant dans la base de données au même moment.

Dans l'autre direction, tous les mots classifiés comme des trous dans la liste d'exclusion pendant le processus de classification, une fois vérifiés dans des corpus, peuvent être ajoutés à OSLIN. Les formes de citation des nouvelles entrées sont stockées sur une liste temporaire, et avec la système de gestion de OSLIN il est facile d'ajouter les mots de cette liste temporaire à MorDebe, avec toutes les formes flexionnelles. De cette façon, la liste d'exclusion n'est pas actualisée avec des formes individuelles, mais avec des paradigmes flexionnels entiers.

Quand on utilise NeoTrack et OSLIN ensemble, la combinaison des deux système résulte dans une base de données lexical, constamment actualisée, et augmenté avec des mots récentes. À cause de cette intégration dans l'ILTEC, MorDebe est un des plus grands lexiques existants pour le portugais.

4. Contexte Théorétique

4. Critère de néologicité et systèmes comparables

NeoTrack utilise une méthode d'extraction de néologismes basée sur l'exclusion, et non pas par exemple, une méthode statistique (comme APRIL) ou bien une méthode basée sur des règles dérivationnels (comme *Zeitgeist*). Grâce à cette méthode d'extraction, il est possible de trouver les mots nouveaux dès leur première apparition. Pour cette raison, les entités capturées par NeoTrack devraient en effet être classifiées comme des *occurrences néologiques* – des cas de « mots » dans un texte pas avant utilisé, pour lesquelles le temps doit encore révéler si ce sont vraiment des néologismes, ou en effet des occasionalismes.

Dans une méthode basée sur l'extraction les néologismes sont déterminés par les mots qui sont exclus. Avec NeoTrack, ceci dépend parfois du contenu de la liste d'exclusion, ainsi que des critères additionnels considérés pendant la phase de classification des candidats à néologismes. Quand NeoTrack est utilisé comme il était pensé, et comme il est utilisé par l'ONP, NeoTrack utilise une combinaison du critère lexicographique (étant donné que la source pour la liste d'exclusion initiale est un dictionnaire) et le critère diachronique (étant donné que des mots sont ajoutés à la liste d'exclusion quand ils sont attestés dans des corpus récents). Cette combinaison s'appelle le *critère lexicographique diachronique*.

Le but de NeoTrack est de compiler tous les néologismes d'un texte en satisfaisant un critère de néologicité déterminé. La raison étant que, pour l'étude de la productivité relative des différents processus de formation de mots nouveaux d'une langue, il est important d'avoir une liste de néologismes impartiale – une liste dans laquelle les mots se trouvent indépendamment de leur fréquence absolue, ou de leur visibilité. Pour une liste complète, il est plus important de présélectionner tous les candidats (rappel) que de supprimer tous les candidats faux (précision), parce qu'il est assez facile d'effacer les candidats faux à la main, mais pratiquement impossible de récupérer une candidat déjà supprimé. À cause de ça, les filtres utilisés dans NeoTrack sont, en comparaison avec par exemple Telenaute, assez légères.

Comme il est expliqué dans le paragraphe 2.3, il est possible d'ajouter des néologismes à NeoTrack manuellement. Les critères utilisés pour des néologismes introduits manuellement seront évidemment différents des critères utilisés pour les néologismes compilés semi automatiquement. Avec la détection manuelle, on ne trouve que des néologismes qui, pour une raison ou une autre ont un « air nouveau » pour le linguiste/lexicographe. En général, cela veut dire que la plupart des mots du lexique potentiel d'une langue qui commencent à être utilisés échappent au processus de détection. De l'autre côté, la détection manuelle normalement permet de dépouiller les néologismes multi-mots et les néologismes sémantiques, qui échappent la détection semi-automatique. Le mélange des néologismes compilés des deux manières, crée certaine confusion quant aux statistiques sur les néologismes. Pour cette raison, dans la base de données de néologismes de NeoTrack, tous les mots possèdent une marque indiquant si le mot a été extrait manuellement ou semi automatiquement. Dans les comptages de fréquences des néologismes, il est possible de limiter les comptages aux néologismes semi-automatiques seulement.

4.2 Néologismes sémantiques et expressions multi-mots

NeoTrack détecte seulement des néologismes d'un certain type : les *néologismes orthographiques* (néologismes dont la séquence de lettres est nouvelle) composées de seulement un mot, qui ne sont pas homographiques avec des mots existants (dans la liste d'exclusion). Ça n'exclut pas seulement les néologismes sémantiques et les nouvelles expressions multi-mots, mais aussi, par exemple, les nouveaux homonymes d'un mot existate, et les mots nouveaux créés par dérivation nulle. Ces restrictions sont assez considérables, mais ne sont pas exclusives de NeoTrack – le même type de restrictions s'applique en principe à tous les systèmes de détection de néologismes basé sur l'exclusion, et ce sont pas des restrictions pouvant être résolues facilement avec ce type de méthode.

La raison principale de cette restriction est que l'extraction est basée seulement sur l'information présente explicitement dans le texte – c'est-à-dire, seulement le mot écrit. Pour ajouter d'autres types d'informations, comme la catégorie grammaticale des mots, leur valeur sémantique, ou l'indication des expressions multi-mots, on a besoin de méthodes statistiques pour calculer cette information pour chaque mot du texte. Bien que le but de cet article ne soit pas de discuter les problèmes de l'utilisation des statistiques dans la détection de néologismes, il est important d'indiquer à quel point ce type de méthode modifiera la recherche de néologisme.

La méthode la plus directe pour utiliser des techniques statistiques pour la détection de néologismes à base de dérivation nulle ou bien des néologismes sémantiques, c'est l'analyse du contexte. Les mots qui apparaissent à côté d'un mot sont des indications significatives de la catégorie grammatical et du sens du mot. Et donc, quand le sens ou la catégorie grammatical du mot change, cette changement doit se révéler comme une modification de fréquence des mots du contexte. Par exemple, quand le verbe anglais *drive* commençait d'être utiliser comme le nom d'un type de stockage informatique, la fréquence avec laquelle on trouve des déterminant comme *the* ou *a* sur la gauche immédiate des occurrences de *drive* augmentais significativement. Et quand le nom *pen* commençait à s'utiliser pour un type de stockage informatique, la fréquence des mots comme *computer* et *capacity* dans la proximité a augmenté en faveur des mots du contexte plus liés avec le sens existant, comme *pencil* ou *paper*.

Bien que cette méthode puisset être efficace, elle ne fonctionne que si le changement de fréquence est assez élevé pour surmonter la variation aléatoire. Et typiquement le changement de fréquences n'est assez remarquable que lorsque l'usage existant est assez peu fréquent dans le discours récent (comme dans le cas du mot *mace* comme arme médiéval) ou lorsque l'usage nouveau est au moins aussi fréquent que le(s) usage(s) existants. Ça veut dire que ce type de méthode détecte plutôt des mots nouveaux déjà bien établis, ce qui est différent de la détection d'occurrences néologistiques realisée avec NeoTrack.

Pour la détection des nouvelles expressions multi-mots, la situation est plus ou moins la même. La méthode la plus usée est basée sur *l'information mutuelle* : des expressions multi-mots peuvent être reconnues dans la mesure où les mots qui font parte d'une expression multi-mots sont utilisés ensemble avec une fréquence plus élevée (que s'attends à base de leur fréquence individuelle). Étant donnée que cette méthode se base sur la fréquence, une fois encore le changement de fréquence doit être assez grand pour être détectable. Par conséquent, cette méthode rend plutôt les nouvelles expressions plus remarquables que les occurrences néologistiques.

Il y a quand même un cas dans lequel une nouvelle expression est détectable dès le début : dans le cas de *multiple schlerosis*, les deux mots n'étaient pas utilisés avant en anglais : cette expression était par conséquent un couple parfait dès son premier usage. Mais les deux parties de cette expression étant nouvelles, il est plus facile de les extraire avec une méthode basée sur l'exclusion, comme NeoTrack, puisque les deux parties de cette expression sont détectables avec NeoTrack pour être des occurrences néologiques.

5. Conclusion

L'objectif principal de cet article a été de montrer que NeoTrack est un logiciel utile et facile à implémenter pour la détection et la gestion de néologismes bien que, comme les autres programmes semi-automatiques, la détection se limite aux néologismes orthographiques. Récemment, des systèmes capables de détecter les néologismes sémantiques ont été mis en place. Néanmoins, ils ne sont pas encore disponibles. Neotrack est construit pour proposer une sélection complète de tous les occurrences néologiques dans une définition spécifique, ce qui permet de disposer d'une description fiable de la distribution des méthodes de formation de mots dans une langue donnée. Depuis cette perspective, NeoTrack offre plus de contrôle pour l'utilisateur (la lexicographe), mais aussi requise de plus de travail manuel.

Un des principaux avantages de Neotrack par rapport aux systèmes du même types est qu'il est disponible (de sources ouvertes) et très facile à implémenter. Cette disponibilité est particulièrement intéressante pour les groupes de recherche réalisant encore le dépouillement de textes manuellement : Neotrack constitue un soutien informatique donnant des options dès le début de l'étude de néologismes, par exemple pour des langues dans lesquelles il n'y a pas encore d'observatoires néologiques. En particulier pour les langues de minoritaires, il représente un avantage considérable puisqu'il fonctionne sans des systèmes de traitement automatique du langage.

Références

Barbosa, Síliva, José Pedro Ferreira & Maarten Janssen. 2008. "MorDebe-Admin. A Lexicon Management System." EURALEX 2008, Barcelone, Juillet 2008.

Cabré, Teresa. 2006. "NEOROM, réseau d'observatoires de la néologie des langues romanes". *Neologica*, vol. 1.

Correia, Margarita; Ana Mineiro, Mafalda Antunes, Maria Doria & Teresa Cabré. 2006. "O Observatório de Neologia do Português Europeu – ONP: criação e apresentaçã". *Dans*: Actas do XX Encontro Nacional da Associação Portuguesa de Linguística (APL). Lisbonne, Portugal.

Isaac, Fabrice. à paraître. "Telanaute: un outil de veille lexicale". CINEO, Barcelona, Mai 2008.

Janssen, Maarten. 2005. « Open Source Lexical Information Network ». Third International Workshop on Generative Approaches to the Lexicon, Geneva, Switzerland, May 2005.

Janssen, Maarten & Tiago Freitas. 2008. "Spock: a spoken corpus klient". Proceeding of LREC 2008, Marrakech, May 2008.

Renouf, Antoinette. 1993. "A Word in Time: first findings from dynamic corpus investigation" *Dans:* Aarts, de Haan and Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam: Rodopi.

Veale, Tony. 2006. "ZeitGeist: A Computational Model of Neologism Processing". *Dans: Proceedings of the Second International Conference of the German Cognitive Linguistics Association*. München, Alemanhe.

Vivaldi, Jordi. 2000. "SEXTAN: prototip d'un sistema d'extracció de neologisms". *Dans:* Cabré, Freixa, &. Solé (eds.) *La Neologia en el tombant de segle*. Barcelona: IULA. pp. 165- 173.