Présentation de l'atelier SemDis 2014 : sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés

Cécile Fabre[1] Nabil Hathout[1] Lydia-Mai Ho-Dac [1] François Morlane-Hondère[1]
Philippe Muller[2] Franck Sajous[1] Ludovic Tanguy[1] Tim Van de Cruys[2]

(1) CLLE-ERSS : CNRS & Université de Toulouse
(2) IRIT-MELODI : CNRS & Université de Toulouse

Résumé. Il s'agit d'un article d'introduction aux actes de SemDis 2014, atelier dédié aux méthodes d'analyse sémantique distributionnelle, avec une focalisation sur la construction de ressources distributionnelles en français. Il décrit les deux tâches qui ont été proposées dans le cadre de l'atelier : la première est une tâche compétitive de substitution lexicale, basée sur le corpus FRWAC. La seconde, plus exploratoire, consiste à analyser un corpus spécifique relevant du champ du TAL. Nous rendons compte de l'évaluation des systèmes qui ont participé à la tâche compétitive, et donnons un aperçu de la diversité des méthodes qui ont été utilisées par les participants dans les deux tâches.

Abstract. This is an introductory paper for the proceedings of the SemDis 2014 workshop, dedicated to distributional semantics methods with a focus on the construction of French distributional resources. We describe the two tasks that have been set up : the first one is competitive. It is a French lexical substitution task, based on the FRWAC corpus. The second one is a more exploratory task, which consists in the analysis of a specific corpus in the NLP field. We report an evaluation of the systems participating in the competitive task, and give a broad overview for both tasks of the diverse methods that have been used by the participants.

Mots-clés : Sémantique distributionnelle, substitution lexicale, tâche partagée, évaluation.

Keywords: Distributional semantics, lexical substitution, shared task, evaluation.

1 Introduction

Les méthodes d'analyse distributionnelle fondées sur le principe harrissien sont aujourd'hui largement répandues. Des expérimentations nombreuses ont été menées, sur différentes langues, et des travaux de synthèse ont permis récemment de stabiliser les notions et les procédures relatives au calcul distributionnel (Baroni & Lenci, 2010; Turney & Pantel, 2010). L'organisation de la première édition de l'atelier SemDis dans le cadre de la conférence TALN, en 2013, visait à rassembler des travaux relevant de cette démarche, avec une focalisation sur les expériences menées sur le français. Il nous a paru en effet utile de faire le point sur le domaine français, initialement marqué par l'importance de travaux précurseurs à la fin des années 1990, qui ont appliqué la méthode distributionnelle au traitement de corpus spécialisés (Bouaud *et al.*, 1997; Habert & Zweigenbaum, 2002) [1], avec des moyens et des objectifs assez éloignés de ceux qui caractérisent aujourd'hui le champ, majoritairement dédié au traitement de très grands corpus de toutes natures.

La deuxième édition de l'atelier SemDis, organisée dans le cadre de TALN 2014, poursuit ce même objectif, en proposant aux participants de prendre part à deux tâches spécifiques :

– Une tâche compétitive de substitution lexicale basée sur des données issues du corpus FRWAC ;
– Une tâche exploratoire sur un corpus spécialisé constitué dans le champ du TAL.

La décision d'organiser deux tâches complémentaires est motivée par l'intérêt de confronter les méthodes distributionnelles à deux contextes nettement différents pour l'interprétation et la validation des relations sémantiques : la première

1. On peut évoquer à ce propos l'organisation d'une journée ATALA en 1999 par B. Habert et A. Nazarenko, intitulée *Approche distributionnelle de l'analyse sémantique*.

tâche offre les moyens de réaliser une évaluation de type extrinsèque des systèmes (Baroni & Lenci, 2011), et passe par l'analyse d'un grand corpus pour faire émerger des fonctionnements sémantiques à large échelle ; la deuxième implique le traitement d'un corpus spécialisé de taille relativement réduite, permettant de mettre au jour l'organisation sémantique d'un domaine clos, sur lequel les participants possèdent une expertise qui facilite l'évaluation intrinsèque des résultats.

Nous décrivons successivement les caractéristiques des deux tâches, tout en présentant brièvement les travaux des 6 participants à l'atelier (3 participations pour chaque tâche).

2 Tâche 1 : substitution lexicale

2.1 Présentation

La première tâche proposée dans cet atelier est une adaptation au français de la tâche SemEval 2007 *Lexical substitution*, telle qu'elle est présentée dans (McCarthy & Navigli, 2009). Étant donné un mot-cible dans une phrase complète, il s'agit de proposer une ou plusieurs unités de substitution qui n'altèrent pas le sens global de l'énoncé. Le choix du substitut est libre. Il est ensuite confronté aux réponses fournies par des annotateurs humains.

Par exemple, si l'on considère le mot *feux* dans la phrase [2] :

*Le policier a été surpris par les **feux** nourris d'un groupuscule terroriste.*

Un substitut envisageable serait *tirs*.

Par contre, dans la phrase :

*On y voit aussi comment sont organisés les pompiers forestiers, qui contrôlent les départs de **feux** de forêts.*

Le mot *incendies* serait plus adapté.

Cette tâche nécessite donc un ensemble d'opérations complexes : non seulement l'identification de mots similaires à la cible (des synonymes, mais pas uniquement) mais aussi la sélection des plus pertinents en fonction du contexte, à la manière des méthodes de désambiguïsation.

Nous avons proposé aux participants d'appliquer une méthode automatique de substitution lexicale à un jeu d'évaluation qui comporte 30 unités lexicales (10 noms, 10 verbes et 10 adjectifs). Pour chaque mot-cible, 10 phrases différentes ont été proposées (soit un total de 300 phrases). Pour chaque phrase, les participants pouvaient proposer jusqu'à 10 mots de substitution, par ordre décroissant de préférence.

Ces phrases ont été sélectionnées dans le corpus FRWAC (voir section 2.2) et nous avons fait appel à des annotateurs humains pour identifier les meilleurs substituts (voir section 2.3). Les soumissions des participants ont donc été évaluées par comparaison avec cette annotation manuelle (voir section 2.4).

2.2 Données

Les 30 mots-cibles du jeu d'évaluation ont été sélectionnés en fonction de leur fréquence (pour garantir l'efficacité de leur analyse distributionnelle), leur polysémie (pour imposer le besoin d'un recours au contexte) et leur substituabilité (pour rendre la tâche accessible aux annotateurs et aux participants). Pour les deux derniers critères, nous nous sommes basés sur les renvois analogiques du Robert présents dans le dictionnaire DicoSyn (Ploux & Victorri, 1998) ci-après RobertSyn.

Au final, les mots sélectionnés vérifient les critères suivants :
– le mot est un nom, adjectif ou verbe présent dans RobertSyn ;
– le lemme du mot a une fréquence supérieure à 500 occurrences dans le corpus FRWAC (Baroni *et al.*, 2009) ;
– le mot est associé à au moins deux sens distincts dans RobertSyn ;
– parmi les synonymes donnés pour chaque sens du mot dans RobertSyn, on trouve au moins deux mots simples (et pas uniquement des locutions) ;
– chaque sens du mot est associé à au moins deux synonymes présentant chacun une fréquence supérieure à 100 occurrences dans le corpus FRWAC.

Le tableau 1 liste les 30 mots-cibles retenus après une sélection manuelle parmi les candidats possibles.

2. Tous les exemples cités dans cet article sont issus du corpus FRWAC et contiennent un mot-cible substituable indiqué en gras.

Noms	Verbes	Adjectifs
affection, capacité, couverture, débit, direction, don, espace, intérêt, montée, vaisseau	*arrêter, commander, entraîner, éplucher, essuyer, faucher, fonder, interpréter, maintenir, taper*	*aisé, compris, grossier, hermétique, incorrect, mince, modeste, obscur, riche, vaseux*

TABLE 1: Les 30 mots-cibles retenus pour la tâche de substitution lexicale

Pour chaque mot-cible, 10 phrases ont été recherchées dans le corpus FRWAC à l'aide du concordancier NoSketch Engine [3] (Rychlý, 2007) afin de représenter sans ambiguïté ses différents sens, sans viser nécessairement un équilibre en nombre d'exemples (voir tableau 2).

sens	n°	phrase
tuer	1	La guerre franco-prussienne **faucha** le jeune artiste à l'âge de 29 ans.
	2	Un psychiatre dont le fils a été **fauché** au front croise un chirurgien qui trie les blessés qu'il opérera et ceux qu'il laissera crever sur place.
renverser	3	Pendant son mandat, un président, conduisant sa propre voiture, **fauche** un piéton et se rend coupable d'un homicide involontaire.
	4	Sur une première offensive italienne, la France récupère le ballon et Zambrotta **fauche** Vieira.
	5	**Fauchée** par une voiture, une promeneuse de 57 ans décède sur le coup, sa belle-soeur est grièvement blessée.
moissonner	6	C'est pourquoi dans les marais, certaines parcelles sont **fauchées** tardivement l'été.
	7	Il y croit, même s'il reste sous le coup d'une condamnation à quatre mois de prison pour avoir **fauché** un champ de maïs transgénique en 2004.
	8	Sa mission : planter (plus de 2 000 arbres), tailler, **faucher**, récolter les fruits, presser les jus pour les propriétaires privés et publics.
voler	9	Louis XV est un mauvais roi parce qu'il s'est laissé **faucher** l'Inde et le Canada par les Anglais.
	10	On picolait un peu - une bouteille d'alcool **fauchée** chez Ceron.

TABLE 2: Les 10 phrases sélectionnées pour représenter les différents sens du mot-cible *faucher*

Le jeu d'évaluation contient ainsi 300 phrases illustrant différents sens des 30 mots-cibles retenus. Les phrases sont nécessairement complètes et bien formées sur le plan syntaxique, aucune correction orthographique ou grammaticale n'a été effectuée. Afin d'éviter les phrases trop longues, certains composants facultatifs situés en début ou fin de phrase ont pu être supprimés, comme dans l'exemple suivant où le composant entre parenthèses a été ôté de la phrase du jeu d'évaluation.

> *C'est pourquoi il se dissimule dans les recoins **obscurs**, guettant le touriste tel la larve de fourmilion ~~(je te rassure, le trou en moins bien sûr...)~~*

De plus, les phrases dans lesquelles le mot-cible apparaissait dans une séquence figée ont été exclues, comme la phrase suivante où le mot-cible *direction* est intégré à la locution *en direction de*.

> *La circulation en **direction** de la Mairie se fera par l'avenue du Maréchal Leclerc.*

Jeu de test. Un jeu de test a été mis à disposition des participants pour la mise au point de leur système. Il s'agit du jeu établi par Van de Cruys *et al.* (2011) et qui concerne 10 noms, avec 10 phrases pour chacun et des substitutions proposées pour chaque phrase. Les 10 noms sélectionnés étaient : *avocat*, *baie*, *carrière*, *feu*, *glace*, *livre*, *pièce*, *reprise*, *timbre*, *voie*. Les phrases étaient également extraites du corpus FRWAC.

2.3 Annotation

L'association de substituts aux mots-cibles pour les 300 phrases du jeu d'évaluation a été réalisée par des annotateurs francophones (étudiants en sciences du langage niveau L3-M2 et chercheurs en linguistique). Chaque phrase a été anno-

3. http://nl.ijs.si/noske/wacs.cgi/first_form

tée par 7 annotateurs différents, chacun pouvant proposer un maximum de 3 substituts. Chaque annotateur avait reçu les consignes suivantes :

Bonjour et merci de participer à la campagne d'annotation SemDis.

Cette annotation correspond à une tâche de substitution lexicale.
30 phrases vont vous être présentées. Chacune comporte un nom, un verbe ou un adjectif écrit en rouge. Votre tâche est de trouver des mots qui peuvent se substituer à ce mot en rouge tout en préservant au maximum le sens de la phrase. Vous pourrez proposer jusqu'à 3 substituts, mais si aucun ne vous vient à l'esprit, n'insistez pas et passez à la phrase suivante.

Exemple de phrase	**Proposition de substitution**
Les trous sont **remplis** de boue.	pleins, gorgés

Les substituts constitués de plusieurs mots sont possibles (ex. 2) mais les mots simples (ex. 1, 3 ou 4) sont à privilégier. Dans la mesure du possible la substitution doit produire une phrase correcte, mais des modifications syntaxiques légères sont tolérées (ex. 3 - changement de préposition, ex. 4 - changement d'ordre des mots).

Exemple de phrase	**Proposition de substitution**
1. J'ai entendu des **tirs**.	*détonations*
2. J'ai entendu des **tirs**.	*coups de feu*
3. Paul a **échoué** dans sa tentative d'assassinat.	*raté*
4. Le **gros** garçon s'amuse comme un fou.	*obèse*

Bonne substitution !

Le recueil des annotations a été réalisé via l'outil de gestion de questionnaires et d'enquêtes en ligne LimeSurvey [4] (voir figure 1).

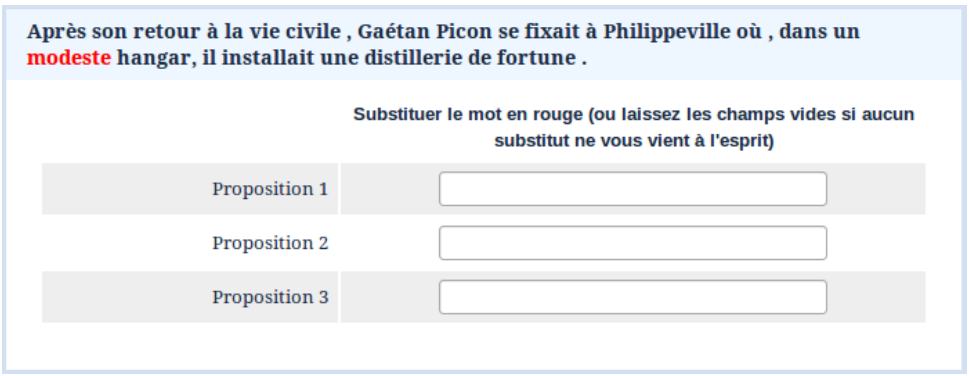

FIGURE 1: Interface d'annotation pour la création du jeu de test

4014 substituts ont été récoltés avec une moyenne de 13 propositions et 7 substituts différents par phrase. Seule une phrase n'a été associée qu'à un substitut : pour la phrase suivante, 3 annotateurs sur 7 ont proposé le mot *peler* comme seul substitut d'*éplucher*, les autres annotateurs n'ayant rien proposé.

> *Olivier Gros, restaurateur, est agacé par le temps mis chaque jour à* ***éplucher*** *et à couper les pommes de terre en diamant (avec des facettes).*

Les données récoltées ont ensuite été nettoyées afin de sélectionner et lemmatiser les substituts du jeu d'évaluation final. Une première validation automatique a permis d'identifier les 3534 propositions qui concernaient exclusivement des substituts mono-lexicaux correctement orthographiés, non ambigus morphologiquement et de même catégorie morpho-syntaxique que le mot-cible. Cette première étape laissait 480 propositions à traiter manuellement.

4. http://www.limesurvey.org

Pour les propositions inconnues d'un lexique du français, une correction orthographique automatique a été appliquée, et le résultat soumis à une validation manuelle. Dans le cas des substituts ambigus, les différents lemmes possibles étaient identifiés et sélectionnés manuellement, comme pour le substitut *prise* (donné pour le mot-cible *faucher*) pour lequel les alternatives *prendre* et *priser* étaient possibles. Les substituts relevant d'une catégorie morpho-syntaxique différente de celle du mot-cible n'ont pas été acceptés, comme par exemple la locution *en tant que* proposée comme substitut du nom *capacité* dans la phrase :

> *Faut-il pour accroître la transparence, que les sessions du Conseil soient publiques, en tout cas lorsque le Conseil agit en sa* ***capacité*** *de législateur ?*

Pour les propositions polylexicales (281 récoltées au total) il a été décidé de supprimer les déterminants, pronoms réfléchis et prépositions périphériques et de conserver les termes jugés « essentiels ». Quelques exemples d'unités polylexicales traitées sont donnés ci-dessous :

> *Dans sa dernière édition, la revue Partir en Croisière consacre sa* ***couverture*** *et son dossier aux Fjords & Glaciers.*
>
> **substitut :** *première page* (proposition initiale : *première page*)
>
> *Encore appelés inhibiteurs calciques, ces médicaments agissent sur les* ***vaisseaux*** *en entraînant leur relâchement*
>
> **substitut :** *canal sanguin* (proposition initiale : *canaux sanguins*)
>
> *J'ai* ***épluché*** *les forums, mais pas de solution à l'horizon, à moins d'investir dans un controlleur RAID onéreux supportant le hot swap.*
>
> **substitut :** *parcourir* (proposition initiale : *parcouru attentivement*)
>
> *90 % de ces hommes ont été* ***arrêtés*** *pour des délits liés à la drogue*
>
> **substitut :** *mettre en examen* (proposition initiale : *mis en examen*)
>
> *Depuis quinze jours, les services de l'urbanisme ont dû* ***éplucher*** *tous les amendements.*
>
> **substitut :** *plonger* (proposition initiale : *se plonger dans*)

Les paraphrases couvrant plus que le seul mot-cible ont été exclues du jeu d'évaluation, comme pour le substitut *se trouvait seule face aux* proposé pour la phrase :

> *En élargissant le débat, un membre du public a remarqué que Wikipédia* ***essuyait*** *quasiment seule les critiques de validation de l'information*

Le bilan du nettoyage est donné dans le tableau 3.

Corrections réalisées	**Nb**
validation automatique	3534
proposition de correction automatique validée manuellement	127
substitut polylexical corrigé manuellement	114
substitut initial conservé	96
substitut initial mal orthographié ou inconnu et corrigé manuellement	81
substitut initial exclu du jeu d'évaluation	53
alternative sélectionnée et validée manuellement	9

TABLE 3: Nettoyage des substituts récoltés

L'accord inter-annotateurs a été calculé sur ces données nettoyées selon les deux mesures utilisées par (McCarthy & Navigli, 2009) :

– **l'accord par paire (*pairwise interannotator agreement*)** mesure la proportion moyenne de réponses identiques pour chaque phrase et pour chaque paire d'annotateurs ;
– **l'accord avec le mode (*mode interannotator agreement*)** mesure la proportion moyenne d'annotateurs qui ont inclus dans leurs réponses le mode, c'est-à-dire la réponse la plus fréquente.

L'accord par paire est de 25,8% et l'accord avec le mode est de 73% [5].

Pour la tâche originale en anglais l'accord par paire mesuré était de 27,75% et l'accord avec le mode de 50,67%. On constate que les taux que nous avons obtenus pour le français sont relativement similaires, avec cependant une très légère baisse au niveau de l'accord par paire et une hausse au niveau de l'accord avec le mode. Ces différences peuvent certainement s'expliquer par le nombre d'annotateurs par phrase : 5 pour la tâche originale contre 7 pour notre tâche.

5. L'accord avec le mode n'est calculé que pour les 77% phrases qui ont un mode.

Le jeu d'évaluation final contient 3961 substituts. Il est disponible librement (sous licence Creative Common) pour des utilisations futures à des fins de recherche [6].

2.4 Évaluation et résultats

2.4.1 Mesures d'évaluation

L'évaluation des soumissions repose sur une comparaison des propositions avec les substituts fournis par les annotateurs. Pour ce faire, nous avons utilisé les mêmes mesures que la tâche SemEval 2007 *Lexical substitution*, à savoir les deux mesures *best* et *oot* (*out of ten*).

– **best** : le système est évalué par rapport à une seule substitution (la meilleure proposition du système, indiquée en premier dans la liste). Le meilleur score est obtenu en proposant le substitut qui est choisi majoritairement par les annotateurs.
– **oot** (*out of ten*) : les soumissions comportent jusqu'à 10 propositions pour chaque mot (sans ordre particulier) et le score calculé correspond au nombre de réponses des annotateurs couvertes par ces propositions. Il n'y a donc aucune pénalité à ajouter des propositions (dans la limite de 10). Ce score permet de mieux prendre en compte la dispersion des réponses des annotateurs.

Pour mieux comprendre les mesures d'évaluation, prenons les annotations d'un exemple du jeu d'évaluation, l'adjectif *mince* (n° 17) :

annotateur (n°)	1	2	3	4	5	6	7
substituts proposés	*étroit*	*étroit*	*étroit*, *fin*	*étroit*, *fin*	*étroit*, *petit*	*fin*, *petit*	*fin*

L'ensemble des réponses agrégées est $H_i = \{$*étroit, étroit, étroit, étroit, étroit, fin, fin, fin, fin, petit, petit*$\}$, et les fréquences associées pour chaque type unique sont $\{$*étroit : 5, fin : 4, petit : 2* $\}$.

Pour calculer le score **best**, on utilise la formule suivante :

$$best(i) = \frac{freq_i(a_i^{best})}{|H_i|} \tag{1}$$

Donc, si un système propose comme meilleur substitut $a_i^{best} =$ *étroit*, il obtient pour cette phrase un score de $\frac{5}{11} = 0,45$. Si le substitut est *petit*, le score est de $\frac{2}{11} = 0,18$. Notons que la valeur maximale possible pour chaque item dépend de la dispersion des réponses des annotateurs.

Pour calculer le score **oot**, on utilise la formule suivante pour évaluer l'ensemble A_i des propositions d'un système pour la phrase numéro i :

$$oot(i) = \frac{\sum_{a \in A_i} freq_i(a)}{|H_i|} \tag{2}$$

Donc, si un système propose comme ensemble de propositions $\{$*fin, petit, épais*$\}$, on obtient pour l'exemple *mince* (n° 17) un score de $\frac{4+2+0}{11} = 0,55$.

Les scores globaux pour un système sont les valeurs moyennes calculées pour les 300 phrases du jeu de test.

Les substituts polylexicaux contenus dans le jeu d'évaluation n'ont pas fait l'objet d'un traitement spécial. Seules les soumissions correspondant parfaitement au substitut ont été considérées comme étant similaires, comme par exemple, la proposition *mettre fin* et le substitut d'évaluation *mettre fin*, mais pas la proposition *canal* et le substitut *canal sanguin*.

Traitement des soumissions Avant d'appliquer les mesures d'évaluation ci-dessus, nous avons traité les soumissions des participants afin de les harmoniser avec les choix effectués lors de l'annotation (voir section précédente). Nous avons donc appliqué les transformations suivantes :

6. http://www.irit.fr/semdis2014/fr/task1.html

– les formes verbales à l'infinitif proposées comme substituts d'un adjectif ont été remplacées par le participe passé. Par exemple, si le système propose *modérer* comme substitut à *modeste*, c'est la forme *modéré* qui sera prise en compte ;
– les verbes pronominaux sont ramenés à la forme principale (si le système a proposé *s'appuyer* comme substitut à *fonder*, c'est la forme *appuyer* qui sera considérée).

Ces traitements ont été faits de façon semi-automatique sur l'ensemble des soumissions avec une vérification manuelle.

Nous avons mis à disposition les données d'évaluation, le script de calcul des scores et les détails des procédures de traitement sur le site Web de la tâche.

2.4.2 Résultats

Nous avons reçu un total de 9 soumissions de 3 participants :

– *Proxteam* (Yann Desalle, Emmanuel Navarro, Yannick Chudy, Pierre Magistry et Bruno Gaume) : 3 soumissions
– *CEA* (Olivier Ferret) : 5 soumissions
– *Alpage* (Kata Gábor) : 1 soumission

Les soumissions de *Proxteam* sont fondées sur des balades aléatoires dans des graphes, qui sont construits à partir de différentes ressources, lexiques (*jeux de mots* et *DicoSyn* pour Proxteam_JDM_Syn) et corpus (*Le Monde* pour Proxteam_LM).

Les soumissions de *CEA* sont fondées sur des modèles de langage neuronaux, où le modèle de neurones estime la probabilité d'un mot en fonction de la séquence de mots qui le précède. Les candidats substituts sont générés à partir de lexiques (*word XP* et *DicoSyn*) et d'un thésaurus distributionnel (*FreDist*).

La soumission d'*Alpage* (*WoDis*) exploite également ces deux types de ressources : la base lexicale *Wolf* est utilisée comme ressource principale, complétée par un thésaurus distributionnel construit à partir de la version française de l'encyclopédie Wikipédia en cas de couverture insuffisante.

Nous avons inclus dans les résultats présentés ici une *baseline* qui utilise la méthode suivante :
– pour chaque mot à substituer, on sélectionne dans le dictionnaire DicoSyn l'ensemble de ses synonymes en ne prenant que les mots simples en compte ;
– ces synonymes sont ordonnés suivant leur fréquence décroissante dans le corpus FRWAC, en limitant les réponses aux 10 premiers synonymes.

Cette *baseline* ne prend aucunement en compte le contexte de la phrase, les réponses sont donc identiques pour toutes les phrases correspondant à un même mot-cible.

La table 4 montre les résultats globaux des systèmes participants et de la *baseline* sur les 300 phrases du jeu d'évaluation. Les soumissions sont classées par ordre décroissant du score *best*.

	best	**oot**
Proxteam_JDM_Syn	.097	.402
CEA_list-word_cos_sent	.075	.236
Proxteam_AxeParaProx_JDM_Syn	.065	.357
Alpage_WoDiS	.063	.205
Proxteam_LM	.051	.212
baseline	.045	.325
CEA_list-fredist_cos_sent	.040	.236
CEA_list-isc_cos_w2	.037	.284
CEA_list-isc_cos_sent	.033	.287
CEA_list-isc_l2_sent	.010	.231

TABLE 4: Résultats globaux des systèmes participants

La table 5 montre les résultats par catégorie grammaticale des systèmes participants (par ordre décroissant du score *best* pour les noms).

	best			oot		
	Nom	*Adj.*	*Verbe*	*Nom*	*Adj.*	*Verbe*
Proxteam_JDM_Syn	.110	.106	.075	.398	.429	.379
CEA_list-word_cos_sent	.075	.074	.076	.195	.245	.268
Proxteam_AxeParaProx_JDM_Syn	.055	.054	.087	.311	.396	.363
Alpage_WoDiS	.054	.072	.061	.191	.211	.213
Proxteam_LM	.052	.040	.061	.233	.166	.237
baseline	.044	.040	.052	.294	.336	.344
CEA_list-fredist_cos_sent	.032	.028	.060	.181	.225	.303
CEA_list-isc_cos_w2	.030	.041	.041	.243	.281	.329
CEA_list-isc_cos_sent	.025	.034	.040	.233	.287	.340
CEA_list-isc_l2_sent	.004	.012	.015	.163	.230	.300

TABLE 5: Résultats par catégorie grammaticale des systèmes participants

Il apparaît donc que c'est le système de l'équipe ProxTeam basé sur des ressources lexicales qui obtient les meilleurs résultats pour cette campagne. On peut toutefois observer des variations importantes de chaque système d'une catégorie grammaticale à l'autre, et bien entendu d'un mot-cible ou phrase à un autre. Ceci nous encourage dans un avenir proche à regarder plus en détails les données récoltées afin de mieux identifier les configurations difficiles et ainsi mieux comprendre les comportements locaux des méthodes appliquées.

3 Tâche 2 : exploration sur le corpus TALN

La deuxième tâche est une tâche exploratoire qui propose aux participants d'examiner plus en détail les résultats de méthodes distributionnelles sur un corpus spécialisé de petite taille.

Pour cela, nous avons proposé aux participants d'utiliser un corpus commun constitué d'une sélection d'articles en français issus des conférences TALN et RECITAL sur la période 2007 à 2013. Il contient environ 2 millions de mots répartis dans 584 articles. Ce corpus est la propriété de l'ATALA ; il a été rassemblé par Florian Boudin (LINA, Université de Nantes) et mis en forme par Ludovic Tanguy (CLLE-ERSS, Université de Toulouse). Pour plus d'information sur le corpus (son origine et son contenu), voir (Boudin, 2013) ; les données elles-mêmes sont disponibles et utilisables librement à des fins de recherche [7].

Nous avons invité les participants à déployer une ou plusieurs techniques d'analyse distributionnelle sur ce corpus, avec les prétraitements et annotations de leur choix. Ceux-ci ont donc pu analyser ce corpus selon leurs objectifs propres, et étudier les phénomènes sémantiques qui leur ont paru les plus pertinents (mise au jour de la polysémie, d'une organisation terminologique, étude de relations sémantiques spécifiques, compositionnalité, etc.). Nous avons cependant demandé, pour illustrer la démarche et les résultats, de privilégier la discussion autour d'un ensemble de mots que nous avons sélectionnés dans le but de faciliter les échanges.

Les mots que nous avons sélectionnés sont les suivants (avec leur fréquence dans le corpus indiquée entre parenthèses) :

1 Verbe	*calculer* (1235)
2 Adjectifs	*complexe* (766), *précis* (376)
5 Noms	*fréquence* (947), *graphe* (1116), *méthode* (3808), *sémantique* (413), *trait* (1806)

Ces mots ont été choisis selon les critères suivants :
- une fréquence minimale pour permettre de déployer confortablement des méthodes distributionnelles classiques ;
- un lien clair avec le domaine du corpus (le TAL) ;
- un potentiel (intuitif) à illustrer un panel de phénomènes sémantiques ; certains mots sont très spécifiques (*graphe*), d'autres ont a priori de nombreux synonymes (*méthode*), d'autres sont polysémiques (*trait*, *précis*), certains ont des acceptions particulières dans ce domaine par rapport à un discours plus général (*trait*, *fréquence*), quant à *sémantique*, il est notoirement difficile à cerner.

7. http://redac.univ-tlse2.fr/corpus/taln.html

Trois équipes se sont prêtées à l'exercice et ont appliqué des méthodes différentes pour examiner le comportement distributionnel de ces 8 mots.

– Ann Bertels et Dirk Speelman ont utilisé une méthode par cooccurrence, en calculant une mesure de similarité basée sur les cooccurrents de deuxième et troisième ordre, et proposé une approche par visualisation des voisins de chacun des 8 mots.
– Gabriel Bernier-Colborne a également utilisé une méthode par cooccurrence (HAL) et sélectionné les voisins les plus proches en faisant varier les différents paramètres impliqués dans le calcul de similarité.
– Cécile Fabre, Nabil Hathout, Franck Sajous et Ludovic Tanguy ont, quant à eux, basé leur approche sur l'exploitation d'une analyse syntaxique, en faisant également varier plusieurs paramètres.

Comme chaque participant a déployé plusieurs configurations (afin notamment d'identifier les paramètres les plus adaptés en fonction du corpus et/ou des mots), c'est en fait une très grande variété d'approches qui ont été appliquées sur ces données. La question de l'interprétation des résultats de ces différentes méthodes est bien entendu cruciale dans l'évaluation ou la comparaison de ces approches. On retrouve là aussi plusieurs façons d'aborder cette question :

– Bertels et Speelman utilisent des représentations graphiques pour pouvoir interpréter les cooccurrents obtenus, et ainsi identifier ceux qui ont un comportement atypique. Pour ce faire ils ont également recours à une observation des contextes correspondants.
– Bernier-Colborne montre (sur un autre corpus) comment des ressources lexico-ontologiques peuvent être utilisées comme données de référence, lorsqu'elles existent (ce qui est le cas pour certains domaines, dont celui de l'environnement qui est utilisé dans l'article). Il montre également comment ces méthodes permettent de dégager des différences et des similarités entre les emplois d'un même mot dans deux domaines différents.
– Fabre et ses collègues ont, quant à eux, eu recours à une annotation manuelle des données pour pouvoir évaluer les différentes méthodes qu'ils ont déployées, et montrent les différences de performance de celles-ci en fonction des mots-cibles et de leur catégorie.

Nous remercions les différents participants pour avoir accepté de jouer le jeu, et espérons que cette première expérience partagée permettra de dégager des principes communs concernant l'utilisation de ces méthodes sur des corpus spécialisés. Il semble en tout cas que le choix d'un domaine spécialisé pour lequel nous possédons tous une compétence interprétative avancée est un atout pour des approches qui, à ce stade, ne peuvent être au final que qualitatives.

Remerciements

Merci à Florian Boudin pour la constitution des archives TALN et au conseil d'administration de l'ATALA qui nous a autorisé à en faire un corpus distribuable et utilisable.

Nous tenons à remercier l'ATILF pour nous avoir permis d'utiliser les dictionnaires qui composent DicoSyn, et notamment les renvois analogiques du dictionnaire *Le Grand Robert* (édition de 1985).

Nous remercions enfin les collègues et étudiants qui ont participé à la campagne d'annotation de la tâche de substitution lexicale.

Références

BARONI M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The WaCky wide web : a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, **43**(3), 209–226.

BARONI M. & LENCI A. (2010). Distributional memory : A general framework for corpus-based semantics. *Computational Linguistics*, **36**(4), 673–721.

BARONI M. & LENCI A. (2011). How we BLESSed distributional semantic evaluation. *Proceedings of the GEMS 2011, Workshop on GEometrical Models of Natural Language Semantics*, p. 1–10.

BOUAUD J., HABERT B., NAZARENKO A. & ZWEIGENBAUM P. (1997). Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles. In *Actes de 1re journées Ingénierie des Connaissances*, p. 207–223, Roskoff.

BOUDIN F. (2013). TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue. In *Actes de TALN*.

HABERT B. & ZWEIGENBAUM P. (2002). Contextual acquisition of information categories. *The Legacy of Zellig Harris : Language and information into the 21st century*, **2**, 203.

MCCARTHY D. & NAVIGLI R. (2009). The English lexical substitution task. *Language Resources and Evaluation*, **43**(2), 139–159.

PLOUX S. & VICTORRI B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement Automatique des Langues*, **39**(1), 161–182.

RYCHLÝ P. (2007). Manatee/bonito - a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, p. 65–70 : Brno : Masaryk University.

TURNEY P. & PANTEL P. (2010). From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research*, **37**(1), 141–188.

VAN DE CRUYS T., POIBEAU T. & KORHONEN A. (2011). Latent vector weighting for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 1012–1022 : Association for Computational Linguistics.