

La désambiguïsation lexicale d'une langue moins bien dotée, l'exemple de l'arabe

Marwa Hadj Salah, Loïc Vial, Hervé Blanchon, Mounir Zrigui, Benjamin Lecouteux, Didier Schwab

▶ **To cite this version:**

Marwa Hadj Salah, Loïc Vial, Hervé Blanchon, Mounir Zrigui, Benjamin Lecouteux, et al.. La désambiguïsation lexicale d'une langue moins bien dotée, l'exemple de l'arabe. 25e conférence sur le Traitement Automatique des Langues Naturelles, May 2018, Rennes, France. hal-01781185

HAL Id: hal-01781185
https://hal.science/hal-01781185

Submitted on 29 Apr 2018

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

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

Traduction automatique de corpus en anglais annotés en sens pour la désambiguïsation lexicale d'une langue moins bien dotée, l'exemple de l'arabe

Marwa Hadj Salah[1, 2] Loïc Vial[1] Hervé Blanchon[1] Mounir Zrigui[2]
Benjamin Lecouteux[1] Didier Schwab[1]
(1) Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP[*], LIG, 38000 Grenoble, France
(2) LaTICE, Tunis, 1008, Tunisie
Prénom.Nom@univ-grenoble-alpes.fr, Prénom.Nom@fsm.rnu.tn

RÉSUMÉ
Les corpus annotés en sens sont des ressources cruciales pour la tâche de désambiguïsation lexicale (*Word Sense Disambiguation*). La plupart des langues n'en possèdent pas ou trop peu pour pouvoir construire des systèmes robustes. Nous nous intéressons ici à la langue arabe et présentons 12 corpus annotés en sens, fabriqués automatiquement à partir de 12 corpus en langue anglaise. Nous évaluons la qualité de nos systèmes de désambiguïsation grâce à un corpus d'évaluation en arabe nouvellement disponible.

ABSTRACT
Automatic Translation of English Sense Annotated Corpora for Word Sense Disambiguation of a Less Well-endowed Language, the Example of Arabic

Sense-annotated corpus are decisive resources for Word Sense Disambiguation (WSD). Most of the languages have none or too little to build robust systems. In this article, we present 12 sense-annotated corpra for the Arabic language automatically build from 12 corpus in English. We evaluate the quality of our WSD systems using a newly available Arabic evaluation corpus.

MOTS-CLÉS : Désambiguïsation lexicale, Construction automatique de corpus annotés, .

KEYWORDS: Word Sense Disambiguation, Automatic translation of annotated corpus, .

1 Introduction

Les corpus annotés en sens sont des ressources cruciales pour la tâche de désambiguïsation lexicale (*Word Sense Disambiguation*). Cette tâche consiste à trouver pour chaque mot d'un texte le sens le plus approprié parmi un inventaire de sens pré-défini. Par exemple, dans la phrase « *Je vois la montagne à travers ma fenêtre.* », l'algorithme devrait choisir le sens du fenêtre qui correspond à la menuiserie plutôt que celui qui correspond à l'interface graphique.

Alors que l'anglais est la langue qui possède la plus grande quantité de telles ressources, la plupart des autres n'en possède pas ou trop peu pour pouvoir construire des systèmes robustes. Nous nous intéressons plus particulièrement ici à la langue arabe. Jusqu'en 2017, et la mise à disposition d'une version étendue de l'*OntoNote 5.0*, aucun corpus annoté manuellement en sens n'était librement disponible. Ce corpus, annoté avec des sens provenant du *Princeton WordNet* anglais pourrait devenir *de facto* le standard d'évaluation de la désambiguïsation lexicale (DL) de l'arabe. Toujours en 2017, notre équipe a également mis à disposition de la communauté la ressource UFSAC qui unifie un

∗. Institute of Engineering Univ. Grenoble Alpes

ensemble de 12 corpus existants en anglais (Vial *et al.*, 2017) annotés avec la version 3.0 du *Princeton WordNet*. Dans cet article, nous adaptons au format UFSAC et étendons la méthode introduite dans (Nasiruddin *et al.*, 2015) et (Hadj Salah *et al.*, 2016). Nous fabriquons ainsi de manière automatique 12 corpus en arabe que nous exploitons pour construire plusieurs systèmes de DL dont nous évaluons la qualité grâce au corpus *OntoNotes Release 5.0.*.

2 Contexte du travail

2.1 Désambiguïsation lexicale

Deux types de ressources sont nécessaires pour la DL : des bases lexicales et des corpus annotés en sens. Ce sont particulièrement les seconds qui sont absents pour la plupart des langues et en particulier pour l'arabe. Trois étapes sont nécessaires pour mettre en place une DL automatique (Schwab, 2017) : 1) *Constitution d'une ressource générique* : plusieurs ressources non dédiées à la DL sont possibles telles que les dictionnaires, les encyclopédies, les corpus non annotés, les corpus annotés, les bases lexicales etc. Certains de ces matériaux sont parfois construits automatiquement en utilisant d'autres matériaux. Cette étape est optionnelle et est souvent réalisée par des équipes spécialisées. 2) *Constitution d'une ressource dédiée à la DL* : on utilise une ou plusieurs ressources brutes pour donner une représentation informatique à chacun des sens d'un mot ; on constitue ici une ressource dédiée à la tâche. Les sens sont soit définis par l'expertise humaine, soit induits à partir des contextes d'utilisation dans les textes (induction de sens). Techniquement, la ressource peut-être un graphe, des définitions ou des représentations vectorielles. 3) *Utilisation de la ressource dédiée pour désambiguïser des textes* : il s'agit de l'algorithme de désambiguïsation proprement dit. Plusieurs facteurs peuvent entrer en compte. Certains sont communs à chaque algorithme comme la taille du contexte considéré pour le mot à désambiguïser (par exemple quelques mots avant ou après celui-ci, la phrase qui le contient, voire le texte) tandis que d'autres dépendent du type d'algorithme mis en œuvre : par exemple la limite à considérer pour la profondeur de la recherche dans un graphe ou encore les paramètres à prendre en compte pour des algorithmes stochastiques.

2.2 Désambiguïsation lexicale de l'arabe

Le fait que les diacritiques soient absents dans les textes arabes est la caractéristique la plus difficile pour la DL, car elle augmente le nombre de sens possibles d'un mot et rend la tâche de désambiguïsation plus difficile. Par ailleurs, la rareté ou la libre disponibilité de ressources (lexicales et/ou annotées) pour l'arabe complique non seulement la création de systèmes de DL pour cette langue mais empêche surtout la comparaison des systèmes entre eux. Pour avancer sur la désambiguïsation de l'arabe, il nous faut donc des corpus annotés en arabe pour apprendre un système de DL ainsi qu'un corpus annoté de référence pour faire de l'évaluation.

3 Méthode mise en oeuvre

Dans cette section nous présentons la méthode mise en œuvre afin de construire automatiquement des corpus annotés en sens dans la langue arabe. Pour ce faire, nous avons besoin de corpus parallèles bilingues afin de construire un système de traduction automatique, un système de DL supervisé, ainsi qu'un corpus de référence annoté en sens pour évaluer la désambiguïsation lexicale.

3.1 Prétraitement du Corpus annoté

Pour traduire un corpus à l'aide de notre système de traduction automatique statistique, celui-ci doit être normalisé pour être dans le même format que les données d'entraînement du système. Pour

ce faire, il est nécessaire d'éliminer les mots composés avec tiret bas existants dans le corpus, les mots non tokenisés, les mots commençant par une majuscule au début d'une phrase, etc. Cette normalisation se fait en trois étapes : 1) segmenter les mots composés (effacer le tiret bas) ; 2) appliquer la tokenisation Moses (ajouter des espaces entre mots et ponctuation) ; 3) mettre chaque mot du corpus dans une balise en suivant le format du corpus et en lui affectant un identifiant unique.

La Figure 1 présente un exemple de normalisation appliquée au mot composé "written_language" :

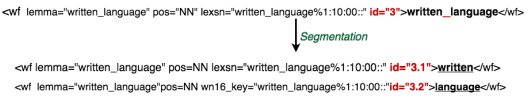


FIGURE 1 – Exemple de normalisation du mot composé "written_language"

3.2 Traduction et portage des annotations

Grâce à la boîte à outils Moses (Koehn *et al.*, 2007) et en exploitant l'ensemble des données parallèles alignées (LDC-Ummah, LDC-News, News Commentary, TED Talks), nous avons construit un système de traduction automatique statistique anglais-arabe afin de traduire de grands corpus annotés en sens et porter les annotations en source vers la langue cible. Nous avons sélectionné au hasard 800 lignes de chacun des corpus (3200 lignes au total) pour les données Test et Dev. Notre système a été évalué avec la métrique BLEU (score de 24,51%). Comme dans nos travaux sur l'amélioration de traduction vers le français d'un corpus annoté dans le contexte de la DL supervisée du français (Hadj Salah *et al.*, 2016) et précédemment dans (Nasiruddin *et al.*, 2015) où nous construisions deux corpus annotés en français et Bengali, nous nous servons des informations d'alignement des mots cible-source fournis par Moses (Koehn *et al.*, 2007) pour transférer les annotations d'un mot source anglais vers son correspondant dans la traduction arabe. (transfert d'annotation d'un mot source vers son correspondant dans la cible).

	LDC-Ummah	LDC-News	News Commentary	TED Talk
Nombre de mots arabes	2M	0.4M	3,9M	0.4M
Nombre de mots anglais	2.4M	0.5M	4.1M	0.5M

TABLE 1 – Description des corpus parallèles utilisés

3.3 Postraitement

Afin d'obtenir les meilleurs résultats possibles, après avoir traduit notre corpus de l'anglais vers l'arabe, nous mettons en œuvre des étapes de post-traitement pour corriger les problèmes (ordre et duplication de mots) posés par l'étape de portage des annotations qui repose sur les informations d'alignement fournies par Moses. Ainsi, nous avons développé un outil pour compiler une chaîne de post-traitement sur la sortie de traduction, qui suit les trois étapes suivantes : 1) Ré-ordonnancement et suppression des mots ajoutés par Moses ; 2) Concaténation des mots composés (ayant les mêmes id) afin d'obtenir un id unique pour chaque mot ; 3) Segmentation permettant d'avoir pour chaque mot, l'information grammaticale correspondante (POS), en utilisant l'analyseur morphologique MADAMIRA. De plus, étant donné que nous traitons de l'arabe comme langue cible, il est nécessaire de passer par une étape de détokensiation de la sortie de traduction pour avoir des mots arabes corrects.

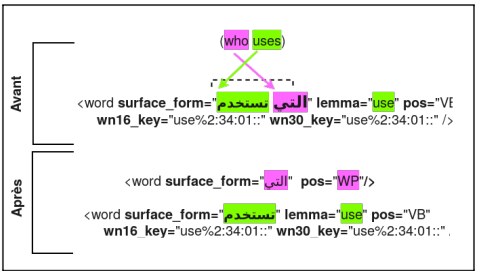


FIGURE 2 – Exemple de post-traitement

4 Application à la désambiguïsation lexicale

4.1 Corpus UFSAC

UFSAC (*Unification of Sense Annotated Corpora and Tools*) est une ressource récemment rendue disponible (Vial *et al.*, 2017). Elle regroupe l'ensemble des corpus en anglais annotés avec une version de *Princeton WordNet*. Ces corpus sont soit directement disponibles lorsque les droits le permettent, soit il est possible de les construire grâce au code source intégré à UFSAC à partir des données originales. UFSAC uniformise également ces corpus avec les sens du *Princeton WordNet* 3.0 (Miller, 1995). Dans les travaux décrits dans cet article, nous exploitons l'ensemble des 12 corpus d'UFSAC (voir tableau 2).

4.2 UFSAC-Arabe

Nous avons appliqué notre méthode pour créer des corpus UFSAC en arabe mais cette méthode pourrait être utilisée pour toute autre langue pourvu que l'on dispose d'un système de traduction de l'anglais vers cette langue.

Ressource	Phrases	Mots		Parties du discours annotées			
		Totaux	Annotés	Noms	Verbes	Adjectifs	Adverbes
SemCor	37176	767415	208142	80552	80079	29977	17534
DSO	101004	2494012	166436	99933	66503	0	0
WNGT	117659	1586199	456880	267985	69886	96299	22710
MASC	31760	548645	102614	45093	35283	11543	10695
OMSTI	820084	28455324	842999	437487	229976	175536	0
Ontonotes	124851	2331961	216283	75354	140929	0	0
SemEval 2007 Task 7	245	5589	1985	997	503	305	180
SemEval 2007 task 17	126	3012	380	135	245	0	0
SemEval 2013 task 12	306	7709	1439	1439	0	0	0
SemEval 2015 task 13	138	2677	959	504	235	144	76
SensEval 2	238	5741	2063	973	492	364	234
SensEval 3	300	5493	1738	806	640	281	11

TABLE 2 – Chiffres relatifs à notre ensemble de corpus en langue arabe annotés en sens

5 Évaluation

Pour évaluer la pertience de notre approche, nous utilisons un système de désambiguisation lexicale supervisée classique en ré-implantant de la méthode NUP-PT utilisée sur l'anglais lors de la compétition SemEval 2007.

5.1 Système de désambiguïsation basé sur les machines à vecteurs de support

L'apprentissage automatique consiste à entraîner un classfieur pour chaque mot cible dans le but de prédire le sens le plus pertinent dans son contexte. Les algorithmes supervisés utilisent des techniques d'apprentissage automatique. Ils apprennent un classifieur sur les corpus annotés en sens en utilisant des classifieurs classiques : séparateurs à vaste marge (NUS-PT (Chan *et al.*, 2007)), classifieurs naïfs bayésiens (NUS-ML, (Cai *et al.*, 2007)), combinaison de séparateurs à vaste marge, entropie maximale (LCC-WSD, (Novischi *et al.*, 2007)). On ne peut pas vraiment affirmer que tel ou tel classifieur soit meilleur qu'un autre et ce qui différencie les performances des systèmes est principalement et directement lié à la taille des données annotées.

Pour cet article, nous avons ré-implanté le classifieur utilisé dans le système NUS-PT qui était le système supervisé état de l'art avant l'émergence des réseaux de neurones profonds. Nous avons fait ce choix pour deux raisons : 1) Prouver la pertience de l'approche ; 2) Utiliser un système de calcul moins gourmand en ressource et donc accessible à un plus grand nombre de chercheurs.

Notre classifieur se base sur trois ensembles de traits pour assigner un sens à un mot donné : 1) Les parties du discours des mots voisins (P_i), 7 traits sont extraits, qui correspondent aux labels de partie du discours des trois mots à gauche (P_{-3}, P_{-2}, P_{-1}) trois mots à droite (P_1, P_2, P_3) et à celui du mot cible (P_0) ; 2) Les collocations locales ($C_{i,j}$), qui correspondent à la suite ordonnée des mots entre les index i et j relativement au mot cible, mis en lettres minuscules. 11 traits sont ainsi extraits : $C_{-1,-1}$, $C_{1,1}$, $C_{-2,-2}$, $C_{2,2}$, $C_{-2,-1}$, $C_{-1,1}$, $C_{1,2}$, $C_{-3,-1}$, $C_{-2,1}$, $C_{-1,2}$ et $C_{1,3}$; 3) Le contexte voisin, ce trait correspond à un vecteur de la taille du nombre de lemmes différents observés pendant l'entraînement. Chaque composante du vecteur correspond ainsi à un lemme, et sa valeur est mise à 1 si le lemme d'un des mots présents dans la même phrase que le mot cible correspond au lemme de cette composante. Elle vaut 0 sinon.

5.2 Corpus d'évaluation : OntoNotes Release 5.0

OntoNotes Release 5.0 (Weischedel *et al.*, 2015) comporte trois langues (anglais, arabe et chinois). C'est un grand corpus annoté manuellement en sens libre de droit contenant plusieurs genres de textes (News, conversations téléphoniques, weblogs, usenet newsgroups, broadcast, talk shows) pour l'anglais et le chinois et seulement des données News pour la partie arabe, avec des informations structurelles (syntaxe et structures prédicat-arguments) et sémantiques superficielles (sens du mot lié à une ontologie et coréférence). La partie arabe d'*OntoNotes Release 5.0* comprend 300K mots du corpus arabe *An-Nahar Newswire*. C'est sur cette partie que nous évaluons notre système tandis que nous l'entraînons en partie sur les traductions de la partie anglaise. Il n'y a pas de biais car les parties arabes et anglaises ne sont pas des traductions l'une de l'autre. Le tableau ci dessous présente le nombre de lemmes, de sens ainsi que les mapping$_{WordNet}$ pour les verbes et noms en arabe.

	#Lemmes	#Lemmes uniques	#Sens uniques	**#Correspondances$_{WordNet}$ uniques**
Verbes	3990	150	642	**4182**
Noms	8534	111	463	**1376**
Total	12524	261	1105	**5558**

TABLE 3 – Description de la partie arabe annotée en sens d'OntoNotes Release 5.0

5.3 Mesures d'évaluation

Nous évaluons notre système de DL *in vitro*, en exploitant le corpus annoté de référence cité précédemment, et en utilisant les mesures d'évaluation classiques comme la plupart des tâches de DL telle

que *SemEval 2013* : en termes de précision P, de rappel R et du score F1 qui correspond à la moyenne harmonique de P et R. La précision se définit comme :

$$P = \frac{\text{nombre de mots annotés correctement}}{\text{nombre de mots annotés}} \quad R = \frac{\text{nombre de mots annotés correctement}}{\text{nombre de mots à annoter}} \quad F1 = \frac{2 \cdot P \cdot R}{P+R}$$

5.4 Résultats et analyse

Dans cette partie, nous présentons les résultats de notre système de DL sur l'anglais afin de montrer qu'il obtient des résultats état de l'art hors réseaux de neurones puis sur l'arabe pour montrer que nous obtenons de bons résultats sur nos corpus automatiquement générés pour l'arabe.

5.4.1 Résultats de DL sur l'anglais

Afin de vérifier l'efficacité de notre méthode et ainsi évaluer l'influence de l'étape de traduction et le portage des annotations sur la qualité de la DL arabe, nous avons entraîné un système de désambiguisation lexicale SVM anglais nommé SVM-UFSAC-eng sur les corpus UFSAC originaux en anglais (Semcor, DSO, WNGT, MASC, OMSTI et OntoNotes anglais) comportant 1.99M mots annotés. Ensuite, ce système a été évalué sur différents corpus ($SensEval2$, $SensEval3task1$ et $semeval2007task17$) pour comparer les résultats.

Système	Tâche	Précision	Rappel	Score F1
SVM-UFSAC-eng	SensEval2	71.34	69.57	70.45
SVM-UFSAC-eng +repli premier sens	SensEval2	71.88	71.88	**71.88**
IMS+emb	SensEval2	68,3	68,3	68,3
SVM-UFSAC-eng	SensEval3task1	65.32	59.58	62.31
SVM-UFSAC-eng+repli premier sens	SensEval3task1	65,50	65.50	**65.50**
IMS+emb	SensEval3task1	68.2	68.2	68.2
SVM-UFSAC-eng	semeval2007task17	60.92	60.65	60.79
SVM-UFSAC-eng+repli premier sens	semeval2007task17	60.87	60.87	**60.87**
IMS+emb	semeval2007task17	68.2	68.2	59.7

TABLE 4 – Performances de notre système de désambiguïsation lexicale sur l'anglais

D'après le tableau ci-dessous, nous remarquons que le système SVM-UFSAC-eng a obtenu les meilleurs performances en termes de score F1 sur les corpus $SensEval2$, $SensEval3task1$ et $semeval2007task17$, respectivement 71.88%, 65.50% et 60.87%, en ajoutant le repli vers le premier sens.

D'autre part, nous avons reporté dans le tableau les résultats du meilleur système de désambiguïsation supervisé état de l'art (Iacobacci *et al.*, 2016) (nommé IMS+emb). Nous observons que les performances de SVM-UFSAC-eng sont comparables.

Dans ce qui suit, nous présentons nos expériences sur la langue arabe, nous montrons que nos résultats sont convenables et encourageants comparés à l'anglais.

5.4.2 Résultats de DL sur l'arabe

Nous avons réalisé l'entraînement de notre algorithme de DL SVM (voir section 5.1) sur les douze corpus traduits depuis l'anglais comportant 2M de mots annotés. Nous appelons ce système SVM-UFSAC-ara.

Le tableau 5 présente les résultats de notre système. Comme beaucoup d'algorithmes de DL, SVM-UFSAC-ara n'annote pas l'ensemble des termes. Par exemple, si les corpus annotés ne contiennent pas d'exemples pour un mot à étiqueter, il ne peut pas réaliser cette opération. Nous utilisons alors l'heuristique classique qui consiste à choisir le premier sens de *Princeton WordNet*.

	Précision	Rappel	Score F1
SVM-UFSAC-ara	68.60	62.14	65.21
SVM-UFSAC-ara+ repli premier sens	67.55	62.74	65.06
SVM-UFSAC-ara+Post-traitement	**70.86**	64.20	**67.36**
SVM-UFSAC-ara+Post-traitement + repli premier sens	69.75	**64.79**	67.18

TABLE 5 – Performances de notre système de désambiguïsation lexicale arabe

Ces résultats montrent qu'il est possible de créer des systèmes de DL pour des langues peu dotées comme l'arabe, qui n'ont pas (ou trop peu) de données annotées en sens. À notre connaissance, il s'agit du premier système évalué sur le corpus OntoNotes, il n'est donc pas possible de se comparer à d'autres systèmes. Toutefois, il convient de noter que nous obtenons des résultats similaires aux résultats obtenus habituellement sur des tâches d'évaluation de l'anglais.

Notre système de désambiguïsation arabe a été évalué en termes de *Précision, Rappel* et *Score F1* sur le corpus de référence OntoNotes arabe. Notre système a réussi à désambiguïser 11346 mots annotés sur 12524 ; les mots non annotés n'étant pas présent dans le corpus d'entraînement.

En lisant le tableau 5, nous pouvons remarquer tout d'abord, qu'en ajoutant le repli vers le premier sens pour les mots non annotés, le système SVM-UFSAC a obtenu 67.55% en termes de précision et 65.06% en termes du score F1. En outre, en appliquant les différentes étapes de post-traitement décrites précédemment sur les données traduites en arabe, notre système de désambiguïsation obtient une meilleure performance en termes de précision 70.86% (+3.31%) , c'est-à-dire qu'il a été capable de désambiguïser correctement 8040 mots parmi les 11346 mots annotés, et en termes du score F1 67.36% (+2.30%).

Par conséquent, on peut dire que nous avons obtenu des résultats de DL arabe similaires aux résultats obtenus pour l'anglais sachant que les corpus d'entraînement utilisés pour les deux langues ont presque la même taille, ce qui prouve l'efficacité de notre méthode.

6 Conclusion et perspectives

Dans cet article, nous avons exposé les difficultés posées par la langue arabe lors de la création et l'évaluation des systèmes de désambiguïsation lexicale. Nous avons montré que l'absence de corpus annotés en sens en est la cause. Pour palier ce manque de ressources, nous proposons à la communauté jusqu'à 12 corpus arabes annotés en sens. Ces corpus sont obtenus par traduction automatique et portage d'annotations de corpus anglais annotés en sens. Ces corpus peuvent par exemple être utilisés pour l'apprentissage de systèmes de désambiguïsation lexicale. Nous avons exploité des corpus annotés pour certaines langues, ici l'anglais, pour créer rapidement un système de désambiguisation lexicale supervisée pour une langue moins dotée telle que l'arabe. Nos résultats prouvent la pertinence de notre approche et sont très encourageants. Il est donc possible de fabriquer des systèmes de désambiguïsation lexicale de bonne qualité pour n'importe quelle langue dès lors que l'on dispose d'un système de traduction automatique de bonne qualité de l'anglais vers cette langue. L'ensemble des douze corpus en arabe et les scripts permettant de les réaliser seront disponibles pour la communauté.

Références

F. BENARMARA, N. HATOUT, P. MULLER & S. OZDOWSKA, Eds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

CAI J. F., LEE W. S. & TEH Y. W. (2007). Nus-ml : Improving word sense disambiguation using topic features. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, p. 249–252, Stroudsburg, PA, USA : Association for Computational Linguistics.

CHAN Y. S., NG H. T. & ZHONG Z. (2007). Nus-pt : exploiting parallel texts for word sense disambiguation in the english all-words tasks. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, p. 253–256 : Association for Computational Linguistics.

G. DIAS, Ed. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

HADJ SALAH M., BLANCHON H., ZRIGUI M. & SCHWAB D. (2016). Amélioration de la traduction automatique d'un corpus annoté. In *JEP-TALN-RECITAL 2016*.

IACOBACCI I., PILEHVAR M. T. & NAVIGLI R. (2016). Embeddings for word sense disambiguation : An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 897–907, Berlin, Germany : Association for Computational Linguistics.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R. *et al.* (2007). Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, p. 177–180 : Association for Computational Linguistics.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benarmara *et al.*, 2007), p. 101–110.

MILLER G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, **38**(11), 39–41.

NASIRUDDIN M., TCHECHMEDJIEV A., BLANCHON H. & SCHWAB D. (2015). Création rapide et efficace d'un système de désambiguïsation lexicale pour une langue peu dotée. In *TALN 2015-22ème Conférence sur le Traitement Automatique des Langues Naturelles,, Caen, France*.

NOVISCHI A., SRIKANTH M. & BENNETT A. (2007). Lcc-wsd : System description for english coarse grained all words task at semeval 2007. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, p. 223–226, Stroudsburg, PA, USA : Association for Computational Linguistics.

SCHWAB D. (2017). Cours master mosig.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benarmara *et al.*, 2007), p. 401–410.

VIAL L., LECOUTEUX B. & SCHWAB D. (2017). Uniformisation de corpus anglais annotés en sens. In *24ème Conférence sur le Traitement Automatique des Langues Naturelles*, Orléans, France.

WEISCHEDEL R., PALMER M., MARCUS M., HOVY E., PRADHAN S., RAMSHAW L., XUE N., TAYLOR A., KAUFMAN J., FRANCHINI M., EL-BACHOUTI M., BELVIN R. & HOUSTON A. (2015). Ontonotes release 5.0. *LDC2013T19. Web Download. Philadelphia : Linguistic Data Consortium*.