Étiquetage morphosyntaxique de langues non dotées à partir de ressources pour une langue étymologiquement proche

Yves Scherrer, Benoît Sagot

Étiquetage morphosyntaxique de langues non dotées à partir de ressources pour une langue étymologiquement proche

Yves Scherrer Benoît Sagot
Alpage, INRIA & Université Paris-Diderot, 75013 Paris
{prenom.nom}@inria.fr

RÉSUMÉ

Nous présentons une approche générique du transfert d'annotations morphosyntaxiques d'une langue dotée vers une langue non dotée étymologiquement proche. Nous ne présupposons aucun corpus parallèle et aucune connaissance préalable de la langue non dotée (ni lexique, ni corpus annoté). Notre approche repose uniquement sur des paires de cognats obtenues par apprentissage non-supervisé selon le paradigme de la traduction automatique statistique à base de caractères, et sur un dictionnaire morphosyntaxique de la langue dotée. Pour les mots fréquents et courts, nous préférons assigner les étiquettes directement aux mots de la langue non dotée en fonction de mesures de similarité inter-langues du contexte morphosyntaxique immédiat. Partant de l'allemand comme langue dotée, nous évaluons notre approche sur le néerlandais, qui est en réalité dotée, et le palatin. Nous obtenons une précision d'étiquetage de 67,2% pour le néerlandais et de 60,7% pour le palatin.

ABSTRACT

Morphosyntactic tagging of non-resourced languages based on resources for another etymologically related language

We introduce a generic approach for transferring part-of-speech annotations from a resourced language to a non-resourced but etymologically close language. We do not rely on the existence of any parallel corpora or any linguistic knowledge for the non-resourced language (no lexicons, no annotated corpora). Our approach only makes use of cognate pairs that are automatically induced in an unsupervised way, based on character-based statistical machine translation and on a morphosyntactic lexicon for the resourced language. Frequent and short words are treated differently, as we tag them directly based on a cross-language similarity assessment of immediate morphosyntactic contexts. Using German as a resourced language, we evaluate our approach on Dutch — in fact a resourced language — and on Palatine German. We reach tagging accuracies of 67.2% on Dutch and 60.7% on Palatine German.

MOTS-CLÉS : étiquetage morphosyntaxique, langues proches, langues non-dotées.

KEYWORDS: part-of-speech tagging, etymologically close languages, non-resourced languages.

1 Introduction

Le traitement automatique des langues régionales pose un certain nombre de difficultés. En général, seulement une quantité limitée de textes écrits est disponible, et ces textes ne sont pas annotés. Mais souvent, il existe une langue mieux dotée en ressources qui est étymologiquement

proche ; en effet, de nombreuses langues régionales d'Europe peuvent être vues — considérations politiques mises à part — comme des dialectes des « grandes » langues standardisées. Notre travail se situe dans ce type de configuration linguistique, où on observe à la fois une asymétrie quantitative des données (peu de données dans la langue régionale, beaucoup de données dans la langue standardisée) et une asymétrie qualitative (données annotées dans la langue standardisée, mais pas dans la langue régionale).

Nous présentons ici une approche générique du transfert d'annotations morphosyntaxiques (parties du discours) d'une langue dotée (ci-après, LD) vers une langue non dotée (LND), où les deux langues sont étymologiquement proches. Ce dernier point nous permet de formuler deux hypothèses quant à la similarité des deux langues. Premièrement, au niveau lexical, les deux langues partagent beaucoup de cognats, c'est-à-dire des paires de mots qui sont formellement (graphémiquement) et sémantiquement similaires. Deuxièmement, au niveau structurel, nous admettons que les systèmes de morphologie flexionnelle des deux langues sont similaires et que l'ensemble des catégories morphosyntaxiques est le même. Ainsi, nous supposons qu'à chaque fois que nous rencontrons une paire de cognats, la ou les catégories associées à l'un des mots peuvent être transférées à l'autre. Pour les mots fréquents et courts, que nous qualifierons abusivement de « mots grammaticaux », cette approche ne donne pas de résultats fiables ; dans ce cas, nous préférons assigner les étiquettes directement aux mots de la langue non dotée selon la similarité du contexte morphosyntaxique immédiat. Le texte ainsi annoté peut ensuite être utilisé comme corpus d'entraînement pour un étiqueteur.

Notre tâche est donc très différente de ce que l'on appelle en général l'*étiquetage morphosyntaxique non supervisé*. En effet, on trouve dans la littérature, depuis Merialdo (1994), de nombreux articles qui définissent sous ce nom la tâche consistant à effectuer un étiquetage morphosyntaxique sans corpus d'entraînement mais grâce à un lexique morphosyntaxique préexistant. Autrement dit, on projette un lexique sur un corpus brut, et la tâche revient alors à désambiguïser l'annotation ainsi obtenue. Dans la plupart des cas, des techniques d'apprentissage automatique sont mises en place afin d'induire un modèle probabiliste (cf. cependant Brill, 1995). Le modèle le plus populaire pour cette tâche est celui des Modèles de Markov Cachés (HMM) (Merialdo, 1994; Goldwater et Griffiths, 2007; Goldberg *et al.*, 2008; Ravi et Knight, 2009), mais d'autres modèles ont également été proposés, et notamment des modèles discriminants (Smith et Eisner, 2005). Différentes techniques d'apprentissage ont été utilisées, et des connaissances linguistiques (en plus du lexique) sont parfois intégrées, notamment pour initialiser les paramètres du modèle HMM.

À l'inverse de ces travaux, nous ne présupposons ici la disponibilité d'aucune ressource pour la LND, mis à part des données textuelles brutes. Nous évaluons notre approche sur des données de différentes langues germaniques. Nous avons choisi l'allemand comme langue dotée, et le néerlandais ainsi que le dialecte palatin [1] comme langues non dotées. La configuration allemand–néerlandais nous a servi pour le paramétrage du modèle, puisque des ressources annotées existent pour le néerlandais. La configuration allemand–palatin est plus proche du champ d'application type de notre méthode, mais nous disposons seulement d'un petit corpus annoté pour l'évaluation.

1. Le palatin (en allemand, *Pfälzisch*), ou francique rhénan du Palatinat, est un dialecte moyen-allemand parlé entre Heidelberg, Mannheim, Kaiserslautern et la frontière française. Il se distingue de l'allemand standard par des différences phonétiques régulières notamment au niveau consonantique, par la simplification de certains paradigmes flexionnels, et par certains phénomènes caractéristiques du langage parlé comme l'élision et la cliticisation. En revanche, les différences lexicales sont limitées.

2 État de l'art

La complexité de notre tâche, due essentiellement à une disponibilité de ressources extrêmement contrainte, nécessite le recours à une combinaison originale de travaux de plusieurs domaines de recherche. Nous recensons ici les travaux desquels nous nous sommes inspirés, ainsi que de certains travaux connexes mais dont nous ne pouvions faire un usage pratique.

2.1 Détection de cognats

Hauer et Kondrak (2011) définissent les cognats comme des mots de langues différentes, mais avec une origine linguistique commune. Deux mots entrant dans une paire de cognats doivent satisfaire les trois propriétés suivantes : ils doivent être phonétiquement ou graphémiquement similaires, ils doivent être sémantiquement similaires, et les correspondances phonétiques ou graphémiques doivent être régulières. Cependant, la plupart des travaux se focalisent sur les deux critères de similarité phonétique ou graphémique.

En général, plus les langues en question sont étymologiquement proches, plus le pourcentage de cognats dans leur vocabulaire est élevé. La détection de cognats est donc une tâche peu coûteuse mais très utile pour notre travail. Un de nos objectifs est de détecter des paires de cognats pour en faire un lexique de traduction, nécessairement bruité. Cette tâche d'induction lexicale a été appliquée par Mann et Yarowsky (2001) : ils évaluent différentes mesures de distance phonétique ou graphémique. Ils distinguent notamment les mesures statiques (indépendantes de la paire de langues) des mesures adaptatives (adaptées à la paire de langues par apprentissage automatique). Sans surprise, les auteurs constatent de meilleures performances avec les mesures adaptatives. Hélas, ces mesures nécessitent un corpus d'apprentissage bilingue dont nous ne disposons pas.

Kondrak et Dorr (2004) présentent une multitude de mesures de distance ou de similarité qui ont l'avantage d'être complètement indépendantes de la langue et donc de ne pas avoir besoin d'un corpus d'apprentissage. Leur tâche est cependant un peu différente : ils veulent mesurer la distance entre des noms de médicaments pour prédire si un nouveau nom de médicament peut prêter à confusion avec des médicaments existants. Parmi les mesures qui opèrent sur des chaînes orthographiques (ils proposent également des mesures pour des chaînes phonétiques, mais qui ne nous intéressent pas ici faute de ressources transcrites phonétiquement), l'algorithme BI-SIM (voir section 3.1.1) donne les meilleurs résultats. Inkpen *et al.* (2005) appliquent ces mesures à la tâche d'identification de cognats dans des langues « proches » (anglais–français), et constatent notamment que des classifieurs supervisés ne font pas mieux que des mesures indépendantes de la langue avec un seuil judicieusement choisi.

2.2 Induction de lexiques à partir de corpus comparables

Les langues régionales d'Europe sont souvent dans une situation de diglossie, et leurs locuteurs maîtrisent presque toujours la langue à grande diffusion environnante. La demande de données traduites, et donc l'existence de corpus parallèles, s'en trouvent fortement réduites. En l'absence de corpus parallèles, on peut naturellement se tourner vers la recherche sur l'induction de lexiques à partir de corpus comparables. Cette ligne de recherche a été inaugurée indépendamment par Fung (1998) et Rapp (1999), dont l'idée consiste à examiner les contextes d'un petit

ensemble de mots-amorces. L'hypothèse est alors qu'un mot français apparaissant souvent dans le contexte du mot *école* sera traduit par un mot anglais qui apparaît souvent à côté du mot *school*. L'inconvénient de cette procédure est le besoin d'une liste de mots-amorces traduits (dans notre exemple, la paire *école–school*), et le besoin de grands corpus pour les deux langues afin de pouvoir construire des vecteurs de similarité suffisamment larges.

Le premier point a été adressé par Fišer et Ljubešić (2011) pour des langues proches : ils utilisent des mots identiques et similaires (selon la mesure BI-SIM) pour construire automatiquement la liste des mots-amorces. De plus, ils disposent de corpus lemmatisés et étiquetés pour les deux langues, ce qui réduit la complexité de la tâche. Malheureusement, cette approche n'est pas adaptée à notre tâche, en raison de la petite taille du corpus et de l'absence d'outils d'analyse pour la LND — notre objectif est précisément de créer ce type d'outils.

Finalement, Koehn et Knight (2002) combinent différentes heuristiques, dont la similarité contextuelle et la similarité formelle des mots, pour induire un lexique de traduction à partir de corpus comparables.

2.3 Traduction automatique statistique à base de caractères

Le principe de la traduction automatique statistique (TAS) consiste à apprendre des alignements entre paires de mots apparaissant ensemble dans un corpus parallèle. Dans la TAS à base de segments, certaines paires de mots sont regroupées en paires de segments, ce qui a permis d'améliorer la performance (Koehn *et al.*, 2003). Récemment, une variation de ce modèle a vu le jour sous la forme de la TAS à base de caractères (Vilar *et al.*, 2007; Tiedemann, 2009). Dans ce paradigme, au lieu d'aligner des mots (et des segments de mots) dans un corpus consistant en phrases, on aligne des lettres (et des segments de lettres) dans un corpus consistant en mots. Evidemment, cette approche fait sens uniquement pour des cognats, où l'alignement des caractères est défini. Elle a donc été utilisée pour la traduction entre langues proches (Vilar *et al.*, 2007; Tiedemann, 2009), ainsi que pour la translitération (Tiedemann et Nabende, 2009). Dans ce dernier papier, les auteurs montrent notamment que la TAS à base de caractères donne de meilleurs résultats que des transducteurs probabilistes tels que ceux proposés par Mann et Yarowsky (2001) pour l'induction lexicale.

Ce modèle nécessite donc comme corpus d'apprentissage un ensemble de paires de mots.[2] Si dans les travaux existants ces données sont extraites de corpus parallèles, nous avons choisi de construire un corpus d'apprentissage de manière automatique en utilisant des mesures de similarité statiques telles que BI-SIM. Nous utiliserons la pipeline standard de la TAS : GIZA++ (Och et Ney, 2003) pour l'alignement des lettres, et de Moses (Koehn *et al.*, 2007) pour l'extraction de segments et le décodage (c'est-à-dire la traduction d'un mot source vers un mot cible).

2.4 Transfert d'annotations morphosyntaxiques

Notre objectif étant de transférer des étiquettes morphosyntaxiques d'une langue dotée vers une langue non dotée, une idée simple serait d'utiliser un corpus parallèle aligné par mots, et

2. On peut également modéliser les frontières des mots, auquel cas on utiliserait des segments de mots ou des phrases courtes comme données d'apprentissage. Cependant, nous nous limitons à des paires de mots simples dans nos expériences.

de copier l'étiquette des mots de la LD vers les mots alignés de la LND. Cette méthode a été proposée par Yarowsky *et al.* (2001) et appliquée à l'étiquetage morphosyntaxique, au chunking de syntagmes nominaux, à l'annotation d'entités nommées et même à l'induction de paradigmes morphologiques. Pour toutes ces tâches, les auteurs observent que les annotations transférées sont très bruitées, notamment à cause des erreurs d'alignement. Ils arrivent à éliminer une partie du bruit en entraînant un étiqueteur sur le corpus transféré et en ré-étiquetant le corpus de la LND. Nous nous inspirons de ce travail, mais l'absence de corpus parallèles nous incite à utiliser des alignements obtenus par similarité graphémique.

Une autre approche de la même problématique a été proposée par Feldman *et al.* (2006). Ils entraînent un étiqueteur sur la LD et l'appliquent, moyennant quelques modifications, à la LND. Évidemment, un tel étiqueteur va rencontrer de nombreux mots de la LND qu'il ne saura pas étiqueter. Feldman *et al.* (2006) poursuivent deux stratégies pour pallier à ce problème. Premièrement, ils utilisent un analyseur morphologique basique pour la LND, écrit à la main, pour prédire les étiquettes potentielles parmi lesquelles l'étiqueteur devra en choisir une. Deuxièmement, ils induisent une liste de cognats à l'aide d'une distance d'édition manuellement adaptée à la paire de langues, qui vont permettre à l'étiqueteur de choisir la catégorie du mot de la LD correspondant. Cette approche est prometteuse, mais n'est pas adaptée à notre contexte pour deux raisons. D'abord, il est difficile de créer un analyseur morphologique, même basique, pour certaines langues ou dialectes non standardisés. Ensuite, selon l'étiqueteur choisi, il peut être difficile de modifier les paramètres internes d'un modèle une fois la phase d'entraînement terminée.

3 Expériences

Comme évoqué dans l'introduction, nous proposons une approche inédite, à notre connaissance, pour transférer des annotations morphosyntaxiques d'une langue dotée (LD) vers une langue non dotée (LND) sans requérir la disponibilité de corpus parallèles. Concrètement, nous partons donc de ressources que l'on peut catégoriser en trois types :

1. Un texte brut en LND qui sera annoté à la fin du processus. Comme évoqué en introduction, nous distinguons dans ce corpus les « mots grammaticaux », définis par convention comme les 200 mots les plus fréquents ayant moins de 6 caractères, des autres mots, dits « mots pleins ». Outre des informations fréquencielles par mot-type, nous extrayons de ce corpus les contextes morphosyntaxiques des mots grammaticaux, en vue d'en prédire la catégorie.
2. Un texte brut de gros volume en LD, dont nous extrayons également la liste des mots-types avec leurs fréquences d'occurrence.
3. Un corpus annoté en parties du discours en LD, utilisé pour extraire un lexique morphologique de base ainsi que les contextes morphosyntaxiques des mots grammaticaux.

Nous appliquons notre méthode aux paires néerlandais–allemand et palatin–allemand. Pour les textes bruts, nous avons choisi des textes de Wikipedia : la Wikipedia entière pour le palatin et l'allemand, et un sous-ensemble de 500 000 mots pour le néerlandais, afin de créer des conditions comparables à l'expérience palatin–allemand. En ce qui concerne le corpus annoté allemand, nous utilisons TIGER (Brants *et al.*, 2002). Ce dernier est annoté avec le jeu d'étiquettes STTS (Stuttgart-Tübingen Tagset) qui comporte 54 étiquettes (Thielen *et al.*, 1999). La table 1 résume la taille de ces données. Pour l'évaluation de notre méthode (cf. section 4), nous utiliserons

Type	Source	Langue	Mots	Mots-types
1	Wikipedia	Palatin	272 111	51 002
1	Wikipedia	Néerlandais	500 151	57 917
2	Wikipedia	Allemand	522 934 377	9 347 129
3	TIGER	Allemand	885 067	85 163

TABLE 1 – Données utilisées

comme référence la Wikipedia néerlandaise annotée à l'aide d'un étiqueteur classique, et un fragment de 100 phrases de la Wikipedia palatine annotée manuellement.

Le modèle que nous proposons sépare le traitement des « mots pleins » de celui des « mots grammaticaux ». Pour les mots pleins, nous essayons de trouver des paires de cognats en LND et LD, et de transférer les étiquettes morphosyntaxiques connues des mots LD vers les mots LND correspondant (sections 3.1 et 3.2). Pour les mots grammaticaux, souvent courts et comportant des correspondances graphémiques irrégulières, la détection de cognats est plus difficile. Au lieu d'induire des paires de cognats, nous essayons donc de prédire leurs étiquettes à l'aide de leurs contextes morphosyntaxiques dans nos ressources textuelles (section 3.3).

3.1 Création du lexique ouvert par alignement de lettres

Pour formaliser la problématique, nous définissons une **paire de cognats catégorisée (PCC)** $(w_{\text{LND}}, w_{\text{LD}}, c)$ comme une entrée lexicale bilingue telle que le mot w_{LND} (de la langue non dotée) et le mot w_{LD} (de la langue dotée) sont des cognats, et que les deux mots ont la catégorie morphosyntaxique c.

Nous créons le lexique de PCC uniquement pour les mots « pleins », et cela en plusieurs phases. D'abord, on récolte, pour chaque w_{LND}, l'ensemble des w_{LD} qui sont le plus similaires, selon une mesure de similarité graphémique indépendante de la paire de langues (section 3.1.1). Ensuite, ces paires de mots sont utilisés pour entraîner un modèle de traduction automatique basé sur les caractères (section 3.1.2). Il s'ensuit une étape de filtrage des résultats ainsi obtenus (section 3.1.3). À ce moment, nous disposons de paires de cognats non catégorisées $(w_{\text{LND}}, w_{\text{LD}})$. Nous rajoutons les informations de catégorie à partir d'un ensemble de paires (w_{LD}, c) extraits du corpus annoté de la langue dotée (section 3.1.5).

Pour des raisons pratiques, nous inférons les paires de cognats dans le sens $w_{\text{LND}} \rightarrow w_{\text{LD}}$. Cela nous permet notamment d'obtenir le même w_{LD} pour plusieurs w_{LND} différents et ainsi de tenir compte de variations orthographiques dans la LND. De telles variations sont moins attendues dans la LD, qui en principe est standardisée. De plus, il est préférable de créer le modèle de langage pour la LD, suivant la distribution quantitative de nos données ; selon l'architecture de la TAS, le modèle de langage doit être situé du côté de la langue cible.

3.1.1 Extraction de paires de cognats potentiels

Nous commençons par créer des listes de mots des deux corpus extraits de Wikipedia. En faisant cela, nous ne tenons pas compte des mots courts ($\leq$ 4 caractères), ni des hapax. De plus,

pour l'allemand, nous nous limitons aux mots apparaissant plus de 1000 fois dans le corpus Wikipedia. [3] Ces filtrages nous donnent 20 225 mots néerlandais, 13 466 mots palatins et 26 951 mots allemands. [4]

La similarité formelle entre deux mots est calculée à l'aide de la mesure BI-SIM (Kondrak et Dorr, 2004). BI-SIM est une mesure de similarité graphémique qui utilise des bigrammes de caractères comme unités de base. Elle ne permet pas d'alignements croisés, et elle est normalisée selon la longueur de la chaîne la plus longue. Ainsi, elle permet de capturer un certain degré de sensitivité au contexte, évite de postuler des alignements non-intuitifs et favorise des associations entre mots de longueur similaire. L'avantage de la mesure BI-SIM est qu'elle est complètement générique et qu'elle ne présuppose donc aucune connaissance des relations étymologiques entre les deux langues. En revanche, elle n'est pas très précise et génère beaucoup d'ambiguïté. Par exemple, les paires *Krais–Kreis* et *Krais–Kraus* obtiennent la même valeur BI-SIM, alors que la première paire, qui est correcte, est phonétiquement plus proche.

Nous gardons pour chaque mot w_{LND} les correspondances $(w_{\mathrm{LND}}, w_{\mathrm{LD}})$ qui maximisent la valeur BI-SIM, mais uniquement à condition que cette dernière se situe au-dessus du seuil de 0,7, choisi empiriquement. Ce seuil nous permet d'éliminer des correspondances qui sont très probablement fausses. Si plusieurs w_{LD} correspondent au même w_{LND}, nous les conservons tous. Il en résulte 10 369 mots néerlandais avec au moins une correspondance allemande, dont 7692 sont non-ambigus (c'est-à-dire qu'ils ont exactement un mot correspondant allemand). Le nombre de mots néerlandais distincts étant de 20 225, on a donc su trouver un correspondant en allemand dans 51,3% des cas. Pour la configuration palatin–allemand, nous obtenons 8437 mots palatins avec des correspondances allemandes (62,7%), dont 6568 non-ambigus.

3.1.2 Entraînement d'un modèle de traduction automatique

La prochaine étape consiste à utiliser les paires de cognats potentiels extraites comme indiqué ci-dessus pour entraîner un modèle de traduction automatique basé sur les caractères ; le but de cette étape est d'une part d'inférer des régularités dans les correspondances graphémiques entre les deux langues, et d'autre part d'obtenir un modèle plus robuste et moins sensible au bruit présent dans les données d'apprentissage que la seule mesure BI-SIM.

Les outils de traduction automatique basés sur GIZA++ et Moses disposent d'un nombre de paramètres élevé. Nous avons testé de nombreuses combinaisons de paramètres sur la paire néerlandais–allemand en fonction du taux d'exactitude sur le lexique d'évaluation présenté dans la section 3.1.4. Faute de place, nous ne pouvons pas décrire toutes ces expériences ici, et nous indiquons donc les paramètres tels que retenus *in fine*.

Modèle de langage Nous avons entraîné un modèle de langage sur le corpus Wikipedia de la langue dotée, c'est-à-dire l'allemand. Nous avons enlevé les mots apparaissant moins de 10 fois dans le corpus afin de limiter la taille et de réduire le bruit. Chaque mot est répété autant de fois qu'il apparaît dans le corpus. Le modèle à 10-grammes (de caractères) a

3. Ce seuil peut paraître très élevé, mais il constitue un bon compromis entre la qualité des résultats et la tractabilité informatique. S'il aurait effectivement été souhaitable d'utiliser le corpus entier, cela s'est avéré intractable puisque chaque mot de la langue source doit être comparé avec chaque mot de la langue cible.

4. Les expériences préalables ont montré qu'une configuration asymétrique, avec un peu plus de mots du côté de la LD, donnait de résultats légèrement supérieurs à une configuration purement symétrique.

obtenu les meilleures performances. Nous avons utilisé IRSTLM (Federico *et al.*, 2008) pour l'entraînement du modèle de langage.

Combinaison d'alignements GIZA++ produit des alignements séparés dans les deux sens. Il existe différentes heuristiques pour combiner ces alignements et en extraire des segments. La combinaison *grow-diag-final* s'est montrée la plus efficace dans nos expériences.

Distortion Dans la terminologie de la TAS, la distortion se réfère à la possibilité de changer l'ordre des éléments. Elle peut être apprise de manière lexicalisée (favorisant le déplacement de certains mots au détriment d'autres), ou non (un poids unique). Nous avons préféré désactiver cette possibilité pour éviter d'apprendre des alignements croisés, que nous supposons très rares dans le contexte de correspondances de mots entre langues proches.

Lissage Nous utilisons le *Good Turing discounting* pour ajuster les poids des alignements rares.

Tuning Moses prévoit la possibilité d'adapter divers poids à l'aide d'un corpus de développement et du *Minimum Error Rate Training*. La taille limitée de nos données a conduit à des résultats peu satisfaisants avec cette approche. Nous avons donc gardé les poids par défaut.

3.1.3 Filtrage des candidats

Le modèle de TAS à base de caractères que nous utilisons permet de générer des candidats allemands qui n'ont jamais été vus tels quels par le modèle de langage. Si ce comportement est souhaitable par exemple pour recombiner des mots composés, il ne l'est pas lorsque les candidats proposés n'existent pas dans la langue allemande. [5] Une première idée pour éliminer ces mauvaises correspondances consistait à générer 50 candidats pour chaque mot de la LND et à en sélectionner le premier qui était un mot attesté dans la Wikipedia allemande. Ce filtre s'est avéré insuffisant : par exemple, le candidat erroné *Frankrijk* "France" était validé parce que ce mot néerlandais apparaissait 21 fois dans la Wikipedia allemande, à côté de 120 000 occurrences du candidat correct *Frankreich*. Nous avons donc choisi un filtre qui repose sur le produit entre le score obtenu par Moses et le score de similarité entre les fréquences des deux mots du candidat dans chacune des deux langues (« score combiné »).

De plus, nous avons constaté que beaucoup de paires induites avec un score faible sont fausses. Considérant qu'il vaut mieux n'obtenir aucune réponse qu'une réponse fausse, nous avons mis en place une stratégie de filtrage complémentaire. Elle consiste à éliminer tous les candidats dont le score combiné est inférieur à 0,5 écarts-type en-dessous de la moyenne de l'ensemble des scores combinés.

3.1.4 Évaluation des paires de cognats potentiels

Pour la paire néerlandais-allemand, nous avons procédé à une évaluation intermédiaire des différentes étapes d'induction lexicale présentées ci-dessus. Nous avons construit à cette fin un dictionnaire bilingue de 26 000 entrées, utilisé comme référence. Il est constitué de paires $(w_{\mathrm{LND}}, w_{\mathrm{LD}})$ extraites à partir des Wiktionnaires (Hanoka et Sagot, 2012) et de la collection OPUS (Tiedemann, 2012). Ces deux sources sont largement complémentaires : les entrées lexicales des Wiktionnaires sont précises, mais couvrent uniquement les formes de base. Pour

5. Notre système a par exemple proposé all. **Bienengarnelen* "crevettes d'abeilles" pour nl. *bijeengeroepen* "convoqué", ou all. **Knuffelfest* "fête de Knuffel" pour nl. *knuffelbeest* "peluche".

	BI-SIM (3.1.1)		Moses (3.1.2)		Filtre fréquence (3.1.3)		Filtre confiance (3.1.3)	
Total	20225		20225		20225		20225	
Sans réponse	9856	(48,7%)	0		0		4040	(20,0%)
Inconnu du lexique	4952	(24,5%)	11061	(54,7%)	11061	(54,7%)	8533	(42,2%)
Correct	2478	(12,3%)	3528	(17,4%)	3938	(19,5%)	3836	(19,0%)
Incorrect	2939	(14,5%)	5636	(27,9%)	5226	(25,8%)	3816	(18,9%)

TABLE 2 – Évaluation des différentes étapes à l'aide du lexique néerlandais

Mot néerlandais	Candidats cognats allemands retenus par BI-SIM (3.1.1)	meilleur résultat Moses (3.1.2)	retenus après filtrages (3.1.3)	Commentaire
vegetatie	vegetative vegetation	vegetation	**vegetation**	désambiguïsation correcte par Moses
groenen	grossen großen groben	groenen	**grünen**	correct, grâce à Moses et au filtrage
amfibieën	—	amphibien	**amphibien**	correct mais inconnu de la référence
enkel	enkel onkel	enkel	**enkel**	faux amis (all. *einfach*)
zweden	zwecken zweigen	zweiten	**zweiten**	faux (all. *schweden*)

TABLE 3 – Exemples d'alignements obtenus pour le néerlandais, à différentes étapes du processus

également pouvoir évaluer les formes fléchies, nous avons choisi d'intégrer les entrées d'OPUS, bien que plus bruités. [6] La table 2 résume les taux de rappel et d'exactitude des différentes étapes, en partant des 20 225 mots-types de notre extrait de la Wikipedia néerlandaise.

Pour l'étape décrite en section 3.1.1, 48,7% des mots néerlandais ne sont associés à aucun candidat allemand dont la mesure BI-SIM dépasse le seuil de 0,7, et 24,5% ne peuvent être évalués directement car ils n'apparaissent pas dans le lexique de référence. Parmi ces derniers se trouvent principalement des noms propres, faciles à traduire puisque souvent identiques dans les deux langues, mais peu informatifs quant aux correspondances graphémiques régulières. Enfin, parmi les paires dont le mot néerlandais est connu, 45,7% (12,3% du total) sont correctes. Après utilisation de Moses (section 3.1.2), nous disposons désormais d'un candidat cognat allemand pour tous les mots néerlandais. Le nombre de paires évaluables passe de 5417 à 9164, sans que les scores de précision ne décroissent trop : 38,5% d'entre elles (soit 17,4% du total) sont correctes. Le filtrage par fréquence permet de faire passer la proportion de paires correctes parmi les paires évaluables à 43,0% (19,5% du total). Le filtrage par confiance réduit à 7652 le nombre de paires évaluables, mais avec une exactitude qui atteint 50,1% (3836 paires).

Quelques exemples concrets de cette évaluation sont donnés dans la table 3.

6. L'alignement des mots d'OPUS a été effectué uniquement à l'aide des propriétés distributionnelles, sans tenir compte des similarités orthographiques. Il est dès lors peu probable qu'une paire de mots, induite de manière erronée par notre algorithme de détection de cognats, ait également été induite par l'algorithme d'alignement d'OPUS.

3.1.5 Transfert d'étiquettes morphosyntaxiques

Les sections 3.1.1 à 3.1.3 ont décrit comment nous obtenons des paires de cognats $(w_{\text{LND}}, w_{\text{LD}})$. Il nous reste à créer un dictionnaire morphosyntaxique de la LD, c'est-à-dire un ensemble de couples de la forme (w_{LD}, c), où c est une étiquette. Dans notre cas, ce dictionnaire est extrait du corpus TIGER ; chaque mot allemand est associé de manière non-ambigüe avec son étiquette la plus fréquente. En fusionnant ces deux listes, nous pouvons finalement créer des paires de cognats catégorisés $(w_{\text{LND}}, w_{\text{LD}}, c)$ qui nous permettent d'annoter une partie du corpus.

Cette procédure peut échouer dans deux situations : premièrement, lorsqu'aucun candidat proposé par Moses passe le filtrage, et deuxièmement, lorsque le candidat proposé par Moses ne se trouve pas dans le dictionnaire morphosyntaxique. Dans ces cas, nous essayons de chercher le mot w_{LND} directement dans le dictionnaire morphosyntaxique de la LD, en comptant sur le fait qu'un certain nombre de mots ont une graphie identique dans les deux langues ; ceci permet notamment de récupérer les étiquettes des symboles de ponctuation, des abréviations, des nombres et des noms propres.

3.2 Extension du lexique des mots pleins par analyse de suffixes

Si un mot plein n'a toujours pas reçu de catégorie syntaxique à ce stade, nous essayons de la retrouver au moyen d'une analyse de suffixes. Nous identifions le plus long suffixe commun au mot non-identifié et à au moins un mot identifié, et nous associons au mot non-identifié l'étiquette associée au mot identifié. Au cas où plusieurs mots identifiés partagent le même suffixe, nous choisissons l'étiquette majoritaire de ces mots. Lorsqu'aucun suffixe du mot inconnu n'a été vu (c'est-à-dire qu'aucun mot ne se termine par le même caractère), nous étiquetons le mot comme *FM* (*foreign material*).

Pour ces deux heuristiques comme pour l'approche décrite au paragraphe suivant, des résultats quantitatifs seront donnés à la section 4.

3.3 Étiquetage des « mots grammaticaux » par similarité contextuelle

Comme indiqué précédemment, l'approche par cognats est peu appropriée pour les « mots grammaticaux ». Nous essayons donc d'inférer les étiquettes morphosyntaxiques pour ces mots à l'aide des contextes morphosyntaxiques immédiats. L'idée est la suivante : le corpus annoté allemand nous apprend que les mots de type *ART* (article) apparaissent souvent entre une préposition et un nom ou entre le début de la phrase et un nom. Si dans le corpus LND semi-annoté, le mot *de* a une distribution similaire, l'étiquette *ART* lui sera associée. La difficulté est que seule une partie du corpus LND est annotée, et ces annotations ne sont pas toujours correctes. Il s'agit donc de rendre cette procédure robuste aux données éparses et bruitées.

Plusieurs contextes syntaxiques peuvent être pertinents ; si le mot immédiatement précédent et le mot immédiatement suivant sont probablement les plus informatifs, on peut également considérer des contextes non-contigus. Nous récoltons les contextes suivants (où W dénote le mot dont on veut déterminer la catégorie, T dénote un mot dont on connaît la catégorie, et X dénote un mot dont on ne connaît pas la catégorie) : TWT, TWXT, TXWT, TW, WT, TXWXT, TWXXT, TXXWT. Considérons par exemple le mot néerlandais *is* "est" (verbe) dans la phrase partiellement

	Moses	Mots identiques	Suffixe	*Foreign Material*	« mots grammaticaux »	**TOTAL**
Palatin — *évaluation sur 100 phrases de la Wikipedia annotées manuellement*						
Couverture	20,0%	18,1%	24,3%	0,9%	36,6%	100%
Précision	71,3%	92,7%	27,8%	16,7%	62,1%	**60,7%**
Néerlandais — *évaluation sur 26 022 phrases de la Wikipedia étiquetées en STTS par MElt (voir texte)*						
Couverture	20,6%	16,9%	19,9%	0,4%	42,2%	100%
Précision	59,2%	86,0%	48,4%	4,8%	73,0%	**67,2%**

TABLE 4 – Résultats de l'évaluation de l'annotation des corpus bruts pour le palatin et le néerlandais à partir des ressources pour l'allemand. Le pourcentage de mots traités par chaque approche est indiqué, ainsi que la précision de chaque approche sur les mots qu'elle a étiquetés.

étiquetée suivante : *het/?? is/?? waarschijnlijk/ADJD dat/?? de/?? trompetsignalen/NN van/?? de/?? vorige/ADJA dag/?? (...)*. Nous en extrayons entre autres le contexte *is ADJD*, de type WT, ou encore le contexte *[début de phrase] X is ADJD*, de type TXWT. Par contre, nous ne pouvons pas extraire de contexte de type TW, puisque l'étiquette du mot précédant *is* est inconnue.

Pour chaque type de contexte, par exemple TXWT, nous représentons alors chaque mot de la LND et chaque catégorie non-ouverte de la LD (ici, l'allemand) par un vecteur dont chaque élément correspond à un contexte particulier et en indique la fréquence : un contexte compte autant de fois qu'il existe de séquences de mots distincts qui lui correspondent. Pour chaque type de contexte et chaque mot grammatical, on classe les catégories candidates selon la similarité décroissante ; nous utilisons la distance euclidienne pour comparer les vecteurs. Les résultats obtenus pour chaque type de contexte sont ensuite moyennés à l'aide de la moyenne harmonique.

À l'issue de cette étape, tous les mots du corpus brut en LND ont été annotés. Il s'agit naturellement d'une annotation perfectible, que l'on pourrait par exemple utiliser comme corpus d'entraînement pour un étiqueteur probabiliste classique. Nous reviendrons sur ce point à la section 5 lorsque nous évoquerons quelques pistes pour des travaux futurs.

4 Résultats et discussion

L'objectif de ce travail est l'étiquetage morphosyntaxique de textes extraits de la Wikipedia palatine et de la Wikipedia néerlandaise, en s'appuyant uniquement sur la proximité étymologique de ces langues avec l'allemand. En plus de l'évaluation ponctuelle de l'induction des paires de cognats néerlandais–allemands (cf. section 3.1), nous présentons ici deux évaluations globales du modèle proposé, une pour la paire néerlandais–allemand, et une pour palatin–allemand.

Pour la première, nous utilisons comme référence la Wikipedia néerlandaise étiquetée par l'étiqueteur classique MElt (Denis et Sagot, 2012). Ce dernier a été entraîné sur le corpus Eindhoven (Uit den Boogaard, 1975; van Grootheest, 1992), dont nous avons converti les étiquettes au standard STTS au préalable. Pour la deuxième, nous utilisons comme référence un sous-ensemble de 100 phrases de la Wikipedia palatine que nous avons annoté manuellement. [7]

7. Cette annotation a été effectuée par le premier auteur, locuteur natif de l'allemand mais pas du dialecte palatin, en

La table 4 montre les résultats de ces deux évaluations, et notamment le pourcentage de mots traités par chacune des approches décrites en section 3 (« Couverture »). On constate que la plus grande partie des mots-occurrences est traitée par l'approche des mots grammaticaux, à cause de la fréquence élevée de ces derniers. Les lignes intitulées « Précision » montrent, indépendemment pour chaque approche, le taux de précision. Les mots identiques obtiennent les meilleurs scores à cause des signes de ponctuation, qui sont très faciles à étiqueter. L'approche Moses et l'approche « mots grammaticaux » obtiennent des résultats comparables. En dernière colonne figure la précision totale de l'étiquetage. Si ces valeurs sont bien en deçà des résultats obtenus habituellement pour l'étiquetage morphosyntaxique, nous les considérons comme très satisfaisants vu la difficulté de la tâche : rappelons que nous ne disposons d'aucune information morphosyntaxique des LND et d'aucun alignement entre mots LND et mots LD.

5 Conclusion et perspectives

Nous avons proposé une méthodologie pour l'annotation en morphosyntaxe d'un corpus brut pour une langue peu dotée à partir de ressources pour une langue étymologiquement proche (lexique morphosyntaxique, corpus annoté, corpus brut volumineux). Cette tâche est plus complexe que la tâche classique d'étiquetage morphosyntaxique non-supervisé, pour laquelle on suppose disponible un lexique morphosyntaxique de la langue peu dotée. Nous avons appliqué cette technique aux paires de langues allemand–palatin et allemand–néerlandais, cette dernière étant destinée à permettre une évaluation plus systématique et un paramétrage de l'approche.

La méthode présentée ici est améliorable de plusieurs façons. Tout d'abord, nous prévoyons de chercher à introduire de l'ambiguïté de façon contrôlée, l'assignation d'une seule étiquette par mot n'étant clairement pas satisfaisante. C'est également une condition pour pouvoir mettre en œuvre de façon utile un étiqueteur morphosyntaxique, qui apprendrait à désambiguïser l'étiquetage en fonction des contextes d'occurrence. Il pourrait ainsi permettre l'étiquetage direct de nouveaux corpus dans la langue non dotée et faciliter l'extraction automatique du lexique morphologique extrait de tels corpus annotés. Par ailleurs, nous avons l'intention d'approfondir la distinction pour l'instant rudimentaire entre mots pleins et « mots grammaticaux ». Par exemple, le mot palatin *gheat* "appartient" est considéré comme un mot grammatical à cause de sa longueur et sa fréquence, et sa catégorie est induite de manière incorrecte. Mais s'il avait été considéré comme mot plein, Moses aurait inféré la correspondance correcte, et il aurait été étiqueté correctement. De plus, nous ne considérons actuellement pas la forme des « mots grammaticaux », bien qu'une partie non-négligeable de ces mots soient identiques dans les deux langues.

Enfin, nous souhaitons valider notre méthodologie sur d'autres paires de langues, qui impliquent d'autres langues dotées que l'allemand, si possible en variant les propriétés typologiques des langues concernées. Nous envisageons ainsi d'appliquer notre approche sur des paires de langues romanes, de langues slaves, mais également sur différents états d'une même langue, par exemple pour induire des ressources adaptées à des textes vieux de quelques siècles à partir de ressources adaptées aux textes contemporains.

Remerciements Ce travail a été financé par le LabEx EFL (ANR/CGI) dans le cadre de l'opération LR2.2.

moins d'une journée de travail.

Références

BRANTS, S., DIPPER, S., HANSEN, S., LEZIUS, W. et SMITH, G. (2002). The TIGER Treebank. *In Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, pages 24–41, Sozopol, Bulgarie.

BRILL, E. (1995). Unsupervised learning of disambiguation rules for part of speech tagging. *In Proceedings of the Third Workshop on Very Large Corpora*, pages 1–13, Cambridge, USA.

DENIS, P. et SAGOT, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, 46(4):721–736.

FEDERICO, M., BERTOLDI, N. et CETTOLO, M. (2008). IRSTLM : an open source toolkit for handling large scale language models. *In Proceedings of Interspeech 2008*, Brisbane, Australie.

FELDMAN, A., HANA, J. et BREW, C. (2006). A cross-language approach to rapid creation of new morpho-syntactically annotated resources. *In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 549–554.

FIŠER, D. et LJUBEŠIĆ, N. (2011). Bilingual lexicon extraction from comparable corpora for closely related languages. *In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'11)*, pages 125–131.

FUNG, P. (1998). A statistical view on bilingual lexicon extraction : from parallel corpora to non-parallel corpora. *Machine Translation and the Information Soup*, pages 1–17.

GOLDBERG, Y., ADLER, M. et ELHADAD, M. (2008). EM can find pretty good HMM POS-taggers (when given a good start). *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 746–754, Columbus, USA.

GOLDWATER, S. et GRIFFITHS, T. L. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 744–751, Prague, République tchèque.

HANOKA, V. et SAGOT, B. (2012). Wordnet extension made simple : A multilingual lexicon-based approach using wiki resources. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie.

HAUER, B. et KONDRAK, G. (2011). Clustering semantically equivalent words into cognate sets in multilingual lists. *In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, pages 865–873, Chiang Mai, Thaïlande.

INKPEN, D., FRUNZA, O. et KONDRAK, G. (2005). Automatic identification of cognates and false friends in French and English. *In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, pages 251–257.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. et HERBST, E. (2007). Moses : Open source toolkit for statistical machine translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07), demonstration session*, Prague, République tchèque.

KOEHN, P. et KNIGHT, K. (2002). Learning a translation lexicon from monolingual corpora. *In Proceedings of the ACL 2002 Workshop on Unsupervised Lexical Acquisition (SIGLEX 2002)*, pages 9–16, Philadelphia, PA.

KOEHN, P., OCH, F. J. et MARCU, D. (2003). Statistical phrase-based translation. *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT'03)*, pages 48–54.

KONDRAK, G. et DORR, B. (2004). Identification of confusable drug names : A new approach and evaluation methodology. *In In Proceedings of COLING 2004*, pages 952–958.

MANN, G. S. et YAROWSKY, D. (2001). Multipath translation lexicon induction via bridge languages. *In Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, pages 151–158, Pittsburgh, PA, USA.

MERIALDO, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–72.

OCH, F. J. et NEY, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

RAPP, R. (1999). Automatic identification of word translations from unrelated English and German corpora. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, Maryland, USA.

RAVI, S. et KNIGHT, K. (2009). Minimized models for unsupervised part-of-speech tagging. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP'09)*, pages 504–512, Singapour.

SMITH, N. et EISNER, J. (2005). Contrastive estimation : Training log-linear models on unlabeled data. *In Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 354–362, Ann Arbor, USA.

THIELEN, C., SCHILLER, A., TEUFEL, S. et STÖCKERT, C. (1999). Guidelines für das Tagging deutscher Textkorpora mit STTS. Rapport technique, Université de Stuttgart et Université de Tübingen.

TIEDEMANN, J. (2009). Character-based PSMT for closely related languages. *In Proceedings of the 13th Conference of the European Association for Machine Translation (EAMT 2009)*, pages 12 – 19, Barcelone.

TIEDEMANN, J. (2012). Parallel data, tools and interfaces in OPUS. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie.

TIEDEMANN, J. et NABENDE, P. (2009). Translating transliterations. *International Journal of Computing and ICT Research*, 3(1):33–41. Special Issue of Selected Papers from the fifth international conference on computing and ICT Research (ICCIR 09), Kampala, Uganda.

UIT DEN BOOGAARD, P., éditeur (1975). *Woordfrequenties in geschreven en gesproken Nederlands*. Oosthoek, Scheltema en Holkema, Utrecht.

van GROOTHEEST, D. (1992). Handleiding bij het Eindhoven Corpus (VU-versie). Rapport technique, Vrije Universiteit Amsterdam.

VILAR, D., PETER, J.-T. et NEY, H. (2007). Can we translate letters ? *In Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39, Prague, République tchèque.

YAROWSKY, D., NGAI, G. et WICENTOWSKI, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. *In Proceedings of the first international conference on Human language technology research*, San Diego, USA.