Analyse syntaxique de langues faiblement dotées à partir de plongements de mots multilingues

Kyungtae Lim, Niko Partanen, Thierry Poibeau

Analyse syntaxique de langues faiblement dotées à partir de plongements de mots multilingues

Application au same du nord et au komi-zyriène

KyungTae Lim — Niko Partanen — Thierry Poibeau

Laboratoire LATTICE
CNRS et École normale supérieure, PSL et Université Sorbonne nouvelle, USPC
1 rue Maurice Arnoux, 92120 Montrouge – France
prenom.nom@ens.fr

RÉSUMÉ. *Cet article présente une tentative pour appliquer des méthodes d'analyse syntaxique performantes, à base de réseaux de neurones récursifs, à des langues pour lesquelles on dispose de très peu de ressources. Nous proposons une méthode originale à base de plongements de mots multilingues obtenus à partir de langues plus ou moins proches typologiquement, afin de déterminer la meilleure combinaison de langues possibles pour l'apprentissage. L'approche a permis d'obtenir des résultats encourageants dans des contextes considérés comme linguistiquement difficiles. Le code source est disponible en ligne (voir https://github.com/jujbob).*

ABSTRACT. *This article presents an attempt to apply efficient parsing methods based on recursive neural networks to languages for which very few resources are available. We propose an original approach based on multilingual word embeddings acquired from different languages so as to determine the best language combination for learning. The approach yields competitive results in contexts considered as linguistically difficult.*

MOTS-CLÉS : *analyse syntaxique, modèles multilingues, plongements de mots, langues peu dotées, same du nord, komi-zyriène*

KEYWORDS: *parsing, multilingual models, word embeddings, low-resource languages, North Saami, Komi-Zyrian*

1. Introduction

Le développement de systèmes automatiques, pouvant analyser avec succès des langues faiblement dotées, est une question cruciale pour le traitement automatique des langues (TAL). La plupart des systèmes d'analyse sont en effet fondés sur des techniques d'apprentissage supervisé nécessitant de grandes quantités de données annotées : la disponibilité de tels corpus est une des conditions principales pour obtenir des performances correctes, quelle que soit la tâche visée. Ce type de techniques est donc bien adapté pour les quelques langues pour lesquelles on dispose de nombreuses ressources en ligne (dictionnaires et surtout corpus annotés), mais l'approche laisse aussi de nombreuses autres langues de côté, du fait de l'absence des ressources nécessaires. Par ailleurs, produire des données annotées en grandes quantités demande beaucoup de moyens (que ce soit au niveau humain ou financier). C'est évidemment un problème majeur pour quantité de langues pour lesquelles ces données sont quasi inexistantes et pour lesquelles on ne dispose pas des moyens nécessaires pour y remédier.

Nous nous intéressons dans cet article au cas de l'analyse syntaxique (cet article reprend en partie la présentation du système développé par le LATTICE pour la tâche d'évaluation CoNLL 2017 (Lim et Poibeau, 2017 ; Lim *et al.*, 2018))[1]. L'analyse syntaxique est une tâche classique, fondamentale pour le TAL et nécessaire pour de nombreuses applications dérivées. Les systèmes d'analyse syntaxique récents les plus performants ou les plus emblématiques du domaine (Weiss *et al.*, 2015 ; Straka *et al.*, 2016 ; Ballesteros *et al.*, 2016), pour n'en citer que quelques-uns, reposent tous sur l'approche décrite dans le paragraphe précédent, c'est-à-dire sur une approche par apprentissage supervisé à partir de grands corpus annotés de la langue visée, souvent l'anglais.

La communauté a toutefois bien conscience qu'il faut aller au-delà des quelques langues bien dotées pour lesquelles on dispose de ressources en masse. D'une part parce qu'il y a des besoins concrets pour d'autres langues : différentes communautés linguistiques, en particulier celles liées à des langues minoritaires ou en danger, ont conscience que l'avenir passe entre autres par l'informatisation des langues et la mise au point d'outils performants, y compris pour le grand public. D'autre part, parce que les langues moins bien dotées posent souvent des questions extrêmement intéressantes sur le plan linguistique, et qui ont été trop longtemps négligées jusqu'ici. Les systèmes entraînés seulement sur l'anglais ne donnent qu'une vision étriquée du TAL, visant l'analyse d'une langue analytique à la morphologie extrêmement pauvre. Au-delà de la quantité de données disponible, la prise en compte de la complexité linguistique, notamment morphosyntaxique, est un autre élément fondamental.

Pour prendre un exemple récent, la tâche d'évaluation lors des conférences *Computational Natural Language Learning* 2017 et 2018 (*CoNLL shared task*) (Zeman,

1. Le code source correspondant aux réalisations présentées dans cet article est intégralement disponible sur le site : `https://github.com/jujbob`.

D. *et al.*, 2017 ; Zeman *et al.*, 2018) portait sur environ cinquante langues (plus précisément quarante-neuf en 2017 et cinquante-sept en 2018), soit à peu près toutes les langues pour lesquelles des données annotées syntaxiquement sont disponibles en quantité significative au format UD (Universal Dependencies) (Nivre *et al.*, 2016). C'est probablement le défi d'analyse syntaxique le plus ambitieux jamais organisé de ce point de vue, mais le chiffre de cinquante langues est à considérer en regard des six mille langues estimées dans le monde. Et même si l'on ne considère que les langues pour lesquelles des données écrites sont disponibles, les cinquante-sept langues de CoNLL 2018 ne permettent de couvrir qu'un échantillon extrêmement restreint de ce qui existe.

En ce qui concerne l'analyse syntaxique automatique, l'approche monolingue et supervisée (c'est-à-dire par apprentissage automatique à partir de corpus annotés représentatifs) est évidemment la plus répandue (l'alternative étant les systèmes reposant sur des grammaires entièrement élaborées à la main). Des chercheurs essaient toutefois depuis un certain temps de concevoir des systèmes multilingues. L'approche multilingue a donné des résultats encourageants aussi bien pour les langues faiblement dotées (Guo *et al.*, 2015 ; Guo *et al.*, 2016) que pour les langues déjà bien dotées, et disposant déjà de ressources comme des dictionnaires et des corpus représentatifs (Ammar *et al.*, 2016a ; Ammar *et al.*, 2016b). Dans ce dernier cas, l'idée est de n'avoir à maintenir qu'un seul modèle d'analyse et de pouvoir l'appliquer ensuite aux différentes langues constitutives du modèle. L'approche multilingue a de plus un avantage, même pour les langues bien dotées : en mettant ensemble plusieurs langues, on peut espérer mieux analyser certains mots ou certaines constructions rares sans dégrader les performances sur des phénomènes plus classiques. Il semble donc y avoir toujours un gain possible.

Ainsi, Ammar et ses collègues (cf. références déjà citées) ont proposé des études portant sur des langues indo-européennes pour lesquelles on dispose déjà de ressources importantes. Ils ont démontré que l'approche *via* un modèle multilingue donne généralement de meilleurs résultats que les modèles monolingues correspondants pour les langues visées.

D'une manière générale, c'est surtout pour les langues moins bien dotées que l'approche multilingue est intéressante. Cette approche peut en fait être mise en œuvre de deux façons différentes. La première consiste à projeter des annotations disponibles d'une langue donnée vers une langue peu dotée *via* un corpus parallèle. Cette approche a été utilisée à plusieurs reprises (notamment dans les références déjà citées (Guo *et al.*, 2015 ; Ammar *et al.*, 2016b)), mais elle nécessite de disposer de données parallèles en quantité suffisante, ce qui est souvent problématique. Le transfert de connaissances nous semble aussi problématique en soi, dans la mesure où cela suppose une relative similarité de structure entre les deux langues visées. Si les deux langues sont trop différentes, l'approche fonctionnera mal, ce qui est insatisfaisant à la fois sur le plan pratique et sur le plan théorique. La seconde approche, celle que nous adoptons ici, vise à produire directement un modèle multilingue, pouvant fonctionner pour plusieurs langues, tout en relâchant les contraintes de structure (voir aussi

Scherrer et Sagot (2014) pour une expérience visant l'étiquetage morphosyntaxique de langues non dotées par transfert depuis une langue dotée, sans utilisation de corpus parallèles).

Dans cet article, nous proposons une approche d'analyse syntaxique utilisant des méthodes à l'état de l'art pour des langues disposant de très peu de ressources structurées (mais pour lesquelles des corpus bruts, c'est-à-dire non annotés, sont disponibles). Notre approche ne nécessite qu'un petit dictionnaire bilingue (ou, *a minima*, une liste élaborée manuellement de mots de la langue visée avec leur traduction dans la langue cible) et l'annotation syntaxique (au format UD) manuelle d'une poignée de phrases de la langue visée. Ces données peuvent être mises au point en quelques heures seulement (moins d'une journée) par une personne connaissant la langue en question. Comme souvent dans ce type de schéma, l'hypothèse que nous faisons est qu'il est possible de transférer des connaissances d'une langue à l'autre entre langues apparentées, mais nous ne faisons pas pour autant l'hypothèse d'une similarité de structure stricte entre les langues. Le point principal est d'identifier des éléments communs au niveau lexical, *via* des plongements de mots (*word embeddings*) multilingues. La source première de comparaison entre une phrase en langue source et une phrase en langue cible est donc lexicale et sémantique, plus que syntaxique (même si la syntaxe joue aussi un rôle primordial, bien évidemment ; c'est d'ailleurs pour cela que les langues choisies pour élaborer le modèle d'analyse doivent être sélectionnées avec attention).

Nous faisons aussi l'hypothèse que les performances dépendent largement des langues utilisées pour élaborer le modèle d'analyse. *A priori*, des langues de même famille et, au sein d'une même famille, des langues étroitement apparentées sont évidemment les meilleurs candidats *a priori*, mais les contacts linguistiques peuvent aussi jouer un rôle. Il existe en effet de nombreux cas de langues où les locuteurs sont tous au moins bilingues et s'expriment le plus souvent dans la langue « dominante », ce qui peut affecter largement leur langue maternelle. Ces phénomènes sont connus, mais relativement peu étudiés, et le TAL peut aider à donner une base statistique et quantitative à l'étude de ces phénomènes d'emprunts et de contagion linguistique. Au cours de l'étude, nous détaillerons plusieurs modèles mettant en jeu des langues génétiquement apparentées et non apparentées, afin de mieux comprendre les limites ou les possibilités de transfert de modèles entre différentes familles de langues.

Afin de mener à bien nos expériences, nous nous penchons sur deux langues finno-ougriennes peu dotées et ayant fait l'objet de peu de recherches en TAL jusqu'ici. Le same du nord[2] est une langue parlée par environ 20 000 personnes au nord de la péninsule scandinave (Suède, Norvège, Finlande). Il existe une dizaine de langues sames (25 000 à 30 000 locuteurs environ au total), mais le same du nord est de loin la langue la plus répandue et celle qui est la mieux supportée par les autorités et les médias (il existe des journaux, ainsi qu'une radio et des émissions de télévision soutenues pu-

2. glottolog.org/resource/languoid/id/nort2671

bliquement). Nous nous intéressons par ailleurs au komi-zyriène[3] (parfois abrégé en komi par la suite), une langue finno-ougrienne de Russie assez éloignée du same. Quasiment tous les locuteurs komis parlent aussi le russe, qui est leur langue essentielle de communication (souvent même entre locuteurs komis). Il y a environ 150 000 locuteurs komis.

Les deux langues (same du nord et komi) sont dans des situations différentes mais possèdent aussi de nombreux points communs pour leur avenir : tous les locuteurs sont bilingues, ils utilisent essentiellement une autre langue de communication (le russe pour les Komis, le finnois, le suédois ou le norvégien pour les Sames du nord) et le komi comme le same du nord étaient dévalorisés jusqu'à récemment. La situation a toutefois changé depuis les années 1980 : les communautés ont pris conscience de l'importance de la préservation de leur langue maternelle, des campagnes de numérisation ont permis de rendre disponibles les écrits existants (la production écrite disponible s'étend sur un siècle environ) et surtout la langue est transmise activement aux enfants, au moins dans certaines régions et dans certaines communautés. Les Sames comme les Komis sont aujourd'hui convaincus de l'importance de développer des outils informatiques pour aider à maintenir et développer leur langue.

Sur le plan informatique, la situation des deux langues n'est pas la même. Le centre d'analyse linguistique de l'université de Tromsø (projet Giellatekno[4]) développe depuis plusieurs années des outils permettant la description des langues finno-ougriennes en général, et du same en particulier. On dispose donc de dictionnaires électroniques assez complets pour le same du nord, incluant les paradigmes de flexion et de conjugaison, ce qui permet d'un côté de générer l'essentiel des formes de la langue et de l'autre, d'analyser dynamiquement des formes linguistiques complexes. Les outils d'analyse (analyse morphosyntaxique et syntaxique) sont en revanche limités, l'approche de l'équipe de Tromsø reposant uniquement sur des automates à nombre fini d'états (éventuellement pondérés), mais sans recours à l'apprentissage automatique. Pour le komi, la situation est beaucoup moins favorable que pour le same du nord. L'équipe de Tromsø a commencé à décrire le komi mais les données disponibles restent relativement embryonnaires.

Il faut enfin noter que, fin 2017, un corpus annoté au format UD a été rendu disponible pour le same du nord. Les données disponibles pour le same sont donc aujourd'hui suffisamment massives pour pouvoir évaluer précisément des analyseurs pour cette langue, mais aussi pour développer des analyseurs de manière traditionnelle, à partir d'un corpus d'entraînement important, comme lors de la campagne CoNLL 2018. Pour le komi, la situation est très différente et il n'existait, à notre connaissance, aucun corpus syntaxiquement annoté pour cette langue au moment où nous avons commencé nos expériences. Les quelques corpus électroniques disponibles sont liés à des travaux réalisés à des fins de documentation linguistique (Blokland *et al.*, 2015 ; Gerstenberger *et al.*, 2016) : ces corpus sont relativement petits et dif-

3. `glottolog.org/resource/languoid/id/komi1268`
4. `http://giellatekno.uit.no/`

ficiles d'utilisation dans une perspective de TAL. Notons enfin que ces langues disposent de données numérisées en quantités relativement importantes, ce qui est utile pour l'élaboration de plongements de mots et permet par ailleurs de compenser en partie le manque de ressources.

L'article est structuré comme suit : nous présentons dans un premier temps l'état de l'art en matière d'analyse syntaxique multilingue (section 2). Nous présentons ensuite le modèle lexical mis au point pour nos expériences (section 3), puis l'architecture du modèle d'analyse à base de réseaux de neurones bidirectionnels, dit BiLSTM (section 4). Nous présentons ensuite le détail des expériences sur le same du nord et le komi-zyriène (section 5), avant de finir par une discussion de ces résultats (section 6), une conclusion et quelques perspectives (section 7). Nous présentons enfin les données mises au point pour le komi-zyriène (embryon de corpus annotés au format Universal Dependencies), ainsi que quelques exemples en annexe à cet article.

2. État de l'art

Depuis les travaux pionniers de Hwa *et al.* (2005), de nombreux groupes se sont intéressés à la mise au point d'analyseurs syntaxiques multilingues, et/ou au transfert de connaissances d'une langue à l'autre, que ce soit dans un cadre d'analyse syntaxique ou pour d'autres tâches, par exemple l'analyse morphosyntaxique. La plupart des méthodes supposent un corpus parallèle, avec des annotations d'un côté (langue source), et non de l'autre (langue cible). La tâche repose alors le plus souvent sur une stratégie de transfert d'étiquettes (c'est-à-dire d'annotations) d'une langue à l'autre, en tenant compte des spécificités de chaque langue. D'autres approches évitent le transfert direct en proposant des stratégies plus ou moins élaborées visant tout d'abord à produire des représentations multilingues avancées, pour éviter les problèmes de transfert d'information. L'apprentissage du parseur est alors réalisé directement sur le modèle enrichi ainsi défini.

Comme on l'a dit, les approches reposant sur la projection d'annotations utilisent un corpus parallèle annoté dans la langue source. Ces annotations sont projetées sur le corpus en langue cible, à partir de quoi un analyseur syntaxique peut être inféré par apprentissage automatique (Smith et Eisner, 2009 ; Zhao *et al.*, 2009 ; Liu *et al.*, 2013). Cette approche est efficace mais elle est principalement confrontée à des problèmes liés à l'alignement des mots lors de l'étape de projection d'annotations. Les méthodes proposées reposent sur des algorithmes de projection robustes prenant en compte un contexte large (Das et Petrov, 2011), ou sur des ressources extérieures comme Wikipédia (Kim *et al.*, 2014) ou WordNet (Khapra *et al.*, 2010), ou bien encore sur la correction *a posteriori* de certaines étiquettes de manière heuristique (Kim *et al.*, 2010).

L'alternative consiste à élaborer directement des modèles d'analyse multilingues grâce aux informations contenues dans des corpus parallèles, ou grâce à des connaissances extérieures, provenant en général de dictionnaires bilingues. L'approche consiste à « apprendre » un modèle d'analyse unique, conjointement pour les deux

langues. Des règles, spécifiées ou non à la main, permettent ensuite d'adapter l'analyse et de tenir compte des spécificités des langues considérées. En dehors de l'analyse syntaxique, les modèles multilingues ont été appliqués à d'autres problèmes de traitement automatique des langues, comme la reconnaissance des entités (Zhuang et Zong, 2010) ou l'analyse des rôles sémantique (Kozhevnikov et Titov, 2012).

D'autres méthodes enfin empruntent aux deux approches précédentes pour créer un modèle d'analyse hybride. Il s'agit alors de produire dans un premier temps une représentation en grande partie indépendante des langues (ou plutôt mêlant les différentes langues dans un seul espace de représentation partagé) puis à « apprendre » un analyseur à partir de cette représentation abstraite et « *crosslingue* » (Täckström *et al.*, 2012). Différents types de ressources peuvent être utilisés dans ce cadre, notamment des corpus parallèles et/ou des dictionnaires bilingues.

Les systèmes plus récents reposent quasi systématiquement sur la notion de plongement de mots (« *word embeddings* » en anglais). Comme précédemment, les systèmes utilisent soit des dictionnaires soit des corpus bilingues, voire des documents parallèles (des légendes d'images ou des pages Wikipédia, par exemple) comme source de connaissances pour inférer un modèle bilingue. Une grande variété d'approches a pu être proposée, mais plusieurs auteurs ont montré que ce sont les données utilisées pour l'apprentissage, plus que l'architecture ou les algorithmes utilisés, qui ont une influence majeure sur le résultat final (Levy *et al.*, 2017 ; Ruder *et al.*, 2017). En gros, à partir des mêmes données, on obtient des résultats très similaires avec des approches en apparence différentes, car dans les faits les algorithmes eux-mêmes sont au final relativement similaires, quel que soit leur point de départ.

L'article de Ruder *et al.* (2017) présente en détail les méthodes fondées sur des représentations lexicales riches. Trois approches sont possibles pour obtenir des plongements de mots bilingues (ou multilingues si on généralise l'approche) : *i*) une première approche consiste à obtenir des représentations sous forme de plongements de mots indépendants pour les deux langues visées (selon la technique introduite par Mikolov *et al.* (2013a) par exemple), puis à mettre en relation les deux représentations obtenues par projection d'un espace sémantique sur l'autre, comme par exemple dans (Artetxe *et al.*, 2016) ; *ii*) élaborer directement un modèle bilingue à partir d'un corpus dans lequel des phrases (voire des documents) des deux langues visées sont déjà en rapport direct (corpus parallèle ou similaire) (Gouws et Søgaard, 2015 ; Gouws *et al.*, 2015) ou *iii*) utiliser un corpus parallèle et un espace sémantique pour chaque langue simultanément (Luong *et al.*, 2015), afin d'obtenir la représentation la plus adéquate en fonction des données fournies en entrée au système.

La mise au point de plongements de mots bilingues et multilingues est un secteur clé de la recherche en TAL à l'heure actuelle. Les tendances visent à réduire les contraintes sur les données en entrée pour obtenir des approches rapides, efficaces et surtout simples à mettre en œuvre. Ainsi, Artexte *et al.* (2017) montrent que quelques dizaines (une cinquantaine environ) de couples de mots bien choisis sont suffisants pour obtenir des plongements de mots bilingues de bonne qualité, au lieu des quelques milliers utilisés dans les expériences précédentes. Une équipe de Facebook a même ré-

cemment montré qu'on pouvait produire des plongements de mots bilingues sans données parallèles ni thésaurus bilingue (Conneau *et al.*, 2017). Cet article a eu un relatif retentissement, mais ses conclusions doivent être nuancées, les résultats n'étant satisfaisants que si les corpus utilisés sont très proches stylistiquement et thématiquement (Vulić *et al.*, 2018).

Dans cet article, nous utiliserons la première méthode qui est facile à mettre en œuvre et qui semble obtenir des résultats très satisfaisants malgré sa simplicité. En ce qui concerne l'architecture du système, nous nous inspirons de Guo *et al.* (2015). La principale différence est que Guo et ses collègues utilisent une approche délexicalisée pour leur analyse, tandis que, conformément au système de Ammar *et al.* (2016a), nous avons recours à des représentations multilingues riches pour l'analyse.

3. Mise au point d'un modèle lexical multilingue

Dans la mesure où les langues que nous souhaitons analyser sont finno-ougriennes, nous nous tournons naturellement vers le finnois pour obtenir des connaissances pertinentes pour l'analyse. Le same du nord a été en contact depuis plusieurs siècles avec le finnois et ce sont surtout deux langues étroitement liées sur le plan génétique (Aikio, 2012, p. 67–69). Le komi est plus éloigné du finnois, mais le finnois reste la langue la plus proche sur le plan linguistique pour laquelle on dispose de ressources importantes. Nous nous sommes également intéressés au russe, sachant que le komi est depuis longtemps en contact avec le russe, et que tous les locuteurs komis sont bilingues (ils parlent aussi russe). On peut donc s'attendre à ce que le russe ait influencé le komi et que ce soit une autre source de connaissances pertinente, les structures copiées du russe étant fréquentes en komi, surtout à l'oral (Leinonen, 2006, p. 241).

Enfin, des expériences avec un corpus anglais seront aussi effectuées : l'anglais n'a pas de lien génétique avec le komi ou le same, ce qui le rend intéressant comme « langue de contrôle » (c'est-à-dire pour comparer les performances obtenues par rapport à des langues de la même famille linguistique, par exemple). Il faut toutefois faire attention aux expériences avec l'anglais : la masse de données disponible pour cette langue permet souvent d'obtenir des résultats relativement corrects malgré tout, la quantité permettant de suppléer partiellement au manque de qualité (ou du moins à l'absence de similarité entre l'anglais et les langues visées lors de l'analyse). L'anglais peut aussi avoir une influence bénéfique en apportant des éléments d'information pertinents pour le niveau lexical-sémantique, ce qui est utile même pour une tâche d'analyse syntaxique.

3.1. *Préparation de ressources linguistiques*

Comme nous l'avons déjà dit, pour les expériences qui suivent nous avons recours aux lexiques bilingues disponibles sur le site Giellatekno[5]. Nous avons par ailleurs utilisé les plongements de mots FastText proposés par Facebook en mai 2017 pour le finnois et le russe (Bojanowski *et al.*, 2017). Il nous faut ensuite générer des plongements de mots similaires pour le same et le komi. Pour ce faire, nous avons en premier lieu recours au corpus Wikipédia, mais il s'agit d'un corpus relativement petit pour les langues visées. Nous le complétons alors avec des corpus disponibles dans le domaine public[6]. Nous produisons enfin les plongements de mots monolingues pour chacune des langues considérées à partir du module FastText de Facebook.

3.2. *Projection de plongements de mots pour obtenir une ressource multilingue*

Dans la section précédente, nous avons décrit comment nous avons obtenu des plongements de mots monolingues pour chaque langue considérée mais, logiquement, chacun de ces plongements a son propre espace vectoriel. Afin d'obtenir des plongements de mots bilingues (voire multilingues, en répétant l'opération plusieurs fois), c'est-à-dire des plongements de mots partageant un espace vectoriel unique, nous utilisons la méthode de transformation linéaire proposée par Artexte *et al.* (2016). Pour effectuer cette transformation, il est nécessaire d'avoir un petit lexique bilingue qui va permettre de définir des « points d'attache » entre les deux espaces vectoriels à mettre en regard. Selon les comparaisons présentées dans Artexte *et al.* (2017, p. 457), la taille des dictionnaires que nous utilisons ici est bien supérieure à ce qui est nécessaire pour effectuer la mise en correspondance des deux espaces vectoriels des langues concernées (tableau 1). Il serait intéressant d'essayer avec de très petits dictionnaires, de quelques dizaines de mots au maximum, afin d'estimer la dégradation des performances dans ce cas de figure, mais comme nous disposons de dictionnaires bilingues contenant plusieurs milliers de mots, nous n'avons pas à ce stade exploré de contextes plus difficiles (mais cela sera nécessaire si l'on doit s'intéresser à d'autres langues ouraliennes, moins bien dotées que le same du nord ou le komi-zyriène).

La méthode de projection des deux espaces sémantiques l'un sur l'autre est la suivante. Soit deux plongements de mots différents, l'un X correspondant à la langue

5. Les dictionnaires pour le same sont disponibles ici : `http://dicts.uit.no/smedicts.eng.html` et les autres dictionnaires sont disponibles à l'adresse suivante : `https://gtsvn.uit.no/langtech/trunk/words/dicts/`. Tous ces dictionnaires sont relativement complets et disponibles sous forme libre, avec une licence GNU GPLv3.

6. Notamment les livres numérisés de la collection Fenno-Ugrica (`https://fennougrica.kansalliskirjasto.fi/`) qui ont été corrigés manuellement par le laboratoire d'appui à la production de ressources électroniques pour les langues régionales de Syktyvkar (`http://komikyv.org/`). Pour le same du nord, nous utilisons le corpus gratuit SIKOR (`http://hdl.handle.net/11509/100`), disponible avec une licence CC-BY 3.0.

Paire linguistique	**Taille du dictionnaire bilingue (nombre de couples de mots)**	**Taille des plongements de mots correspondants**
Finnois-same du nord	12 398	2,4 Go
Same du nord-finnois	10 541	2,4 Go
Same du nord-anglais	1 499	1,4 Go
Finnois-komi	12 879	2,3 Go
Komi-anglais	8 746	7,5 Go
Russe-komi	12 354	5,7 Go

Tableau 1. *Taille des dictionnaires et des plongements de mots liés générés à partir des différents dictionnaires (il s'agit de dictionnaires de formes fléchies, ce qui explique que la taille du dictionnaire finnois-same du nord soit par exemple différente de celle du dictionnaire same du nord-finnois).*

cible, et l'autre Y à la langue source, et soit $D=\{(x_i,y_i)\}_{i=1}^{m}$ (*où* $x_i \in X$, $y_i \in Y$) la ressource obtenue consistant en une collection de plongements de mots bilingues. Le but est, dès lors, de trouver la matrice de transformation W telle que xW soit la meilleure approximation de y. On obtient ce résultat en minimisant la somme des carrés des erreurs, suivant Mikolov *et al.* (2013b) :

$$\arg\min_{W} \sum_{i=1}^{m} \|x_i W - y_i\|^2 \qquad [1]$$

Une dégradation importante des résultats peut se produire si la transformation linéaire est appliquée à deux plongements de mots sans autre contrainte. Pour répondre à ce problème, Artetxe *et al.* (2016) proposent une méthode de correspondance orthogonale qui permet de garder un niveau de performance correct. C'est cette variante de l'algorithme que nous avons utilisée ici.

3.3. *Corpus annotés au format Universal Dependencies*

Nous avons également besoin de corpus annotés pour nos expériences, au moins pour montrer leur apport quand ils sont disponibles. Nous avons utilisé des corpus pour l'anglais, le finnois et le russe : tous provenaient de l'initiative Universal Dependencies et peuvent être trouvés en ligne[7].

7. Sur le projet Universal dependencies, voir `http://universaldependencies.org`. Nous avons utilisé les corpus arobés suivants, dans leur version 2.1 : `https://github.com/UniversalDependencies/UD_English-EWT` (anglais), `https://github.com/UniversalDependencies/UD_Russian-GSD` (russe) et `https://github.com/UniversalDependencies/UD_Finnish-TDT` (finnois).

4. Modèle d'analyse en dépendances *crosslingue*

Les analyseurs syntaxiques traditionnels emploient des méthodes d'apprentissage supervisé fondées sur des séries de traits définis en grande partie manuellement. Le développeur doit en fait définir des combinaisons pertinentes (*feature functions*) de traits et de relations entre ceux-ci, afin que le système soit capable de déterminer les relations entre têtes et dépendants[8]. La définition manuelle de ces combinaisons de traits est une tâche difficile et en grande partie arbitraire, que tous les concepteurs de systèmes cherchent à contourner.

Les systèmes récents à base de réseaux de neurones ont plutôt recours à des méthodes automatiques permettant de simplifier le problème, en laissant le soin à la machine de déterminer les combinaisons de traits pertinentes. Ainsi, Chen et Manning (2014) ont proposé d'utiliser des classificateurs non linéaires intégrés dans un modèle de réseau neuronal. Avec cette méthode, les caractéristiques lexicales et non lexicales sont encodées dans des vecteurs qui peuvent être concaténés pour alimenter un classificateur non linéaire. Cette approche présente deux avantages : *i*) les classificateurs non linéaires ont globalement de meilleures performances que les classifieurs linéaires pour identifier les relations entre les éléments pertinents pour l'analyse, et *ii*) cette approche réduit drastiquement le travail manuel dans la mesure où le réseau de neurones se fonde essentiellement sur les caractéristiques calculées par les classifieurs.

4.1. *Architecture du système d'analyse*

Notre approche est ici similaire à celle de Chen et Manning (2014) et de Kiperwasser et Goldberg (2016) pour la partie analyse, mais nous utilisons des plongements de mots multilingues, alors que nos prédécesseurs s'en tiennent à un système monolingue.

Représentations LSTM bidirectionnelles. Les progrès récents en TAL sont largement dus à des représentations sous forme de traits portant des informations efficaces pour l'analyse des relations entre les mots de la phrase (Cho, 2015 ; Huang *et al.*, 2015). Une représentation LSTM bidirectionnelle (bi-LSTM) est un type de réseau de neurones récurrent, où chaque élément dans la séquence à analyser est lui-même représenté par un vecteur. L'algorithme procède en produisant des représentations préfixes (dites *forward* car la phrase est analysée de gauche à droite) et des représentations suffixes (dites *backward* car la phrase est alors analysée de droite à gauche). Un item est représenté par la concaténation de ses deux contextes, gauche

8. "*Traditionally, state-of-the-art parsers rely on linear models over hand-crafted feature functions. The feature functions look at core components (e.g. "word on top of stack", "leftmost child of the second-to- top word on the stack", "distance between the head and the modifier words"), and are comprised of several templates, where each template instantiates a binary indicator function over a conjunction of core elements (resulting in features of the form "word on top of stack is X and leftmost child is Y and ...*")." (Kiperwasser et Goldberg, 2016).

et droit. Soit par exemple la phrase $t = (t_1,t_2,...,t_n)$, dans laquelle le symbole ∘ dénote une opération de concaténation. La fonction LSTM bidirectionnelle correspond à : $\text{BiLSTM}(t_{1:n}, i) = \text{LSTM}_{Forward}(t_{1:i}) \circ \text{LSTM}_{Backward}(t_{i:n})$.

L'architecture est exactement la même que celle du BIST-parser (Kiperwasser et Goldberg, 2016). Nous renvoyons donc le lecteur à cet article fondateur pour connaître les détails de l'architecture du système qui est de ce point de vue relativement standard. Nous avons juste étendu cet analyseur de manière à le rendre multilingue, ce qui oblige à prendre en compte des représentations contextuelles construites par le module bi-LSTM multilingue (un code pour chaque mot, dit *one hot encoding*, permet de déterminer la langue associée).

Représentation lexicale. Soit une phrase en entrée $t = (t_1,t_2,...,t_n)$, une forme lexicale w, une étiquette morphosyntaxique correspondante p, un plongement de mots obtenu préalablement xw et une valeur de codage de la langue concernée l, un mot t_i (*token*) est défini comme : $t_i = e(w_i) \circ e(p_i) \circ e(xw_i) \circ e(l_i)$, où e réfère au plongement de chaque trait et $e(xw_i)$ est le plongement de mots déjà présenté en section 3. Nous ajoutons un code pour désigner la langue concernée, comme dit précédemment (Naseem *et al.*, 2012 ; Ammar *et al.*, 2016a). La plupart des analyseurs monolingues utilisent des traits comme $e(w_i)$ et $e(p_i)$, ainsi que d'autres éléments comme la distance entre la tête et le dépendant, ou d'autres traits spécifiques calculables à partir du corpus UD. Notez enfin que t_i de $\text{BiLSTM}(t_{1:n}, i)$ permet de stocker les contextes *forward* et *backward* du LSTM.

4.2. *Modèle d'analyse*

Il existe deux approches principales en matière d'analyse syntaxique en dépendance. La première est fondée sur la notion de transition (Nivre, 2004), l'autre sur la notion de graphe (McDonald *et al.*, 2005b). Nous utilisons ici une approche à base de graphes héritée du BIST-parser. L'approche semble efficace pour les corpus au format Universal Dependencies, et on renverra à Dozat *et al.* (2017) pour une comparaison détaillée et argumentée des deux approches.

À partir des représentations des mots et de leurs annotations dans la couche BiLSTM, le BIST-parser produit un arbre candidat pour chaque couple dépendant-mot-tête. Les scores attachés aux différents arbres candidats sont ensuite calculés à l'aide d'un perceptron multicouche (MLP), utilisé comme simple fonction de pondération (*scoring function*). Enfin, le système choisit les meilleurs arbres d'analyse en dépendance sur la base de la somme des scores attachés aux différents sous-arbres. Pour plus d'informations sur le modèle à base de graphes et sur le modèle de pondération d'arcs utilisé par le BIST-parser, voir (Taskar *et al.*, 2005) et (McDonald *et al.*, 2005a).

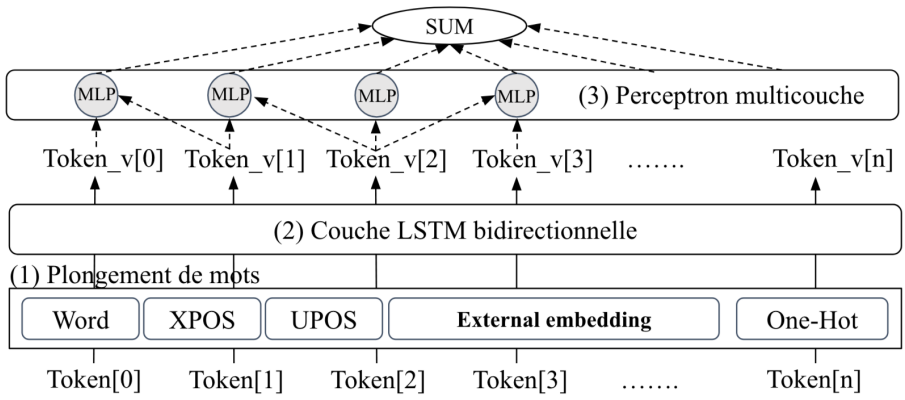


Figure 1. *Architecture du réseau de neurones*

5. Expériences

Nous présentons dans cette section les expériences que nous avons menées sur le same du nord et sur le komi-zyriène.

Corpus disponibles. Le same était une des langues dites *surprise language* de la campagne d'évaluation CoNLL 2017 : seulement vingt phrases étaient fournies pour l'entraînement et les participants n'avaient que quelques jours pour produire un système opérationnel. Le komi n'était pas inclus dans l'évaluation CoNLL 2017 mais nous avons choisi cette langue pour des raisons linguistiques (il s'agit d'une langue finno-ougrienne, comme le same) et parce qu'elle correspond à un cas typique de langue sous-dotée, comme on l'a vu dans l'introduction. Pour pouvoir mener à bien nos expériences, nous avons produit un corpus annoté composé de dix phrases komies pour l'entraînement et soixante-quinze phrases pour le test (ce corpus contient aujourd'hui près de trois cents phrases et grossit régulièrement, mais moins d'une centaine de phrases étaient disponibles au moment des expériences rapportées ici). Une présentation du corpus komi et des problèmes d'annotation rencontrés est disponible à la fin de cet article, en annexe.

Étiquettes morphosyntaxiques utilisées. Dans les expériences rapportées ici, nous nous fondons sur les étiquettes (catégories morphosyntaxiques) fournies par UDpipe (Straka *et al.*, 2016) pour le same et, en l'absence d'analyseur morphosyntaxique disponible pour le komi, nous utilisons les étiquettes posées à la main pour cette langue (*gold*). Lors de la campagne CoNLL 2017, l'analyse se fondait aussi sur des étiquettes de référence pour le same du nord, mais nous préférons recourir ici à un analyseur morphologique pour le same afin de rendre les conditions expérimentales plus proches de la réalité. Nous n'utilisons pas de traits morphologiques autres que ceux du corpus de référence ou ceux fournis par UDpipe pour le same.

Équipe	Score LAS	Score UAS
C2L2 (Ithaca)	48,96	58,85
IMS (Stuttgart)	40,67	51,56
HIT-SCIR (Harbin)	38,91	52,51
Notre système	**28,39**	**42,72**

Tableau 2. *Meilleurs résultats (officiels) pour le same lors de la tâche commune CoNLL 2017 et résultat obtenu par le LATTICE lors de cette même évaluation*

Le fait d'utiliser des étiquettes morphosyntaxiques de référence est bien évidemment quelque peu artificiel, mais permet de se focaliser uniquement sur le niveau syntaxique. Il n'en reste pas moins que la production d'analyseurs morphosyntaxiques performants est évidemment une condition nécessaire pour produire des analyses en situation réelle. La tâche commune CoNLL 2018 impliquait de développer une chaîne complète allant du texte brut à l'analyse syntaxique, et l'expérience a montré que les systèmes conservent ainsi des performances satisfaisantes. On renverra donc le lecteur aux actes de la tâche commune CoNLL 2018 sur ce point (Zeman *et al.*, 2018).

Conditions d'entraînement du système. Comme nous voulons explorer des scénarios pour des langues faiblement dotées, nous avons supposé que l'on ne disposait pas de données de développement permettant d'ajuster les paramètres du système (même dans le cas du same, pour lequel il existe maintenant des données importantes, notamment le corpus annoté syntaxiquement au format UD). Nous avons donc limité les expériences, notamment lors de la phase d'apprentissage, en considérant toutes les données disponibles une fois, sans arrêt anticipé, suivant Guo *et al.* (2016). D'autres stratégies seraient possibles (plusieurs itérations en faisant varier les phrases utilisées lors de l'apprentissage par exemple), mais le gain observé sur les résultats est minime et souvent non significatif. Ce type d'approches pose en outre des problèmes de réplicabilité et nous l'avons donc laissé de côté. Enfin, il faut noter que, pour l'élaboration d'un modèle multilingue, les différentes sources de données sont de taille très déséquilibrée. Pour pallier ce problème, et suivant les travaux antérieurs de Guo *et al.* (2016), nous avons effectué vingt fois plus d'itérations pour les langues faiblement dotées que pour les autres langues.

Comparaison avec la tâche commune CoNLL 2017. Nous avons utilisé les mêmes conditions pour l'entraînement de notre système dans les expériences décrites ici que pour la tâche commune CoNLL. En particulier nous n'avons pas de corpus de développement (nous disposons juste de vingt phrases annotées pour le same et de dix phrases annotées pour le komi pour la mise au point du système, comme dit plus haut).

Le tableau 2 présente les résultats obtenus par les trois meilleures équipes sur le same lors de la tâche commune CoNLL 2017, ainsi que les résultats de notre propre système. Il est évident, au vu de ces résultats, que notre système était alors loin d'être aussi performant que les meilleurs systèmes sur le same, à savoir ceux de Cornell (Shi *et al.*, 2017) , de Stuttgart (Björkelund *et al.*, 2017) ou de Harbin (Che *et al.*, 2017).

Lors de CoNLL 2017, C2L2 (Cornell Univ.) a obtenu les meilleures performances pour le same avec une approche par transfert délexicalisé (en utilisant un corpus de finnois pour l'entraînement et un corpus de développement de vingt phrases en same pour ajuster les paramètres du système, sans utilisation de traits lexicaux, c'est-à-dire en se fondant uniquement sur les étiquettes morphosyntaxiques et non sur les mots eux-mêmes). IMS (Stuttgart) a utilisé une approche similaire (approche par transfert délexicalisé) en utilisant pour l'entraînement quarante corpus différents encodés au format UD, et a ainsi obtenu le deuxième meilleur résultat.

Comparaison avec la tâche commune CoNLL 2018. Le same était à nouveau une langue de test lors de la campagne d'évaluation CoNLL 2018. Le meilleur système a obtenu les résultats suivants lors de la campagne 2018 : 69,87 LAS et 76,66 UAS. Ces résultats sont bien meilleurs que ceux rapportés pour l'évaluation CoNLL 2017 (tableau 2) ou même dans cet article (tableau 3), mais en 2018 des données d'entraînement importantes étaient fournies pour le same (il s'agissait essentiellement du corpus de same au format UD publié après la campagne d'évaluation 2017, comme indiqué dans l'introduction). Il est donc important de souligner que les résultats obtenus sur le corpus CoNLL 2018 ne sont en rien comparables avec les résultats 2017, où seules vingt phrases étaient disponibles pour la mise au point des systèmes.

Les résultats du meilleur système lors de la campagne d'évaluation CoNLL 2018 (69,87 LAS et 76,66 UAS) donnent malgré tout une idée de l'état de l'art pour une langue à morphologie riche, comme le same, quand on dispose d'un corpus d'entraînement moyennement volumineux. Ils permettent aussi d'avoir une idée de l'écart de performance entre une langue pour laquelle on dispose de données annotées et une langue pour laquelle on ne dispose pas de telles ressources (entre 10 et 15 points de différence environ), et aussi une idée de l'écart par rapport à l'anglais (là aussi, entre 10 et 15 points de différence – ces chiffres sont évidemment à prendre avec précaution car il faudrait faire d'autres expériences, avec d'autres langues et des conditions expérimentales plus directement comparables pour obtenir des résultats vraiment fiables). On confirme là, par l'observation, des résultats très évidents : une langue synthétique à morphologie riche est plus complexe à analyser qu'une langue analytique avec une complexité morphologique moindre, et un corpus d'entraînement de grande taille est aussi un facteur majeur d'amélioration des performances. Pour le reste, redisons-le : les résultats CoNLL 2018 ne sont en rien comparables aux résultats 2017 du fait des conditions expérimentales radicalement différentes pour le same lors de ces deux campagnes.

6. Résultats et analyse

Les résultats pour le same du nord sont donnés dans le tableau 3, et les résultats pour le komi-zyriène dans le tableau 4. Les résultats du tableau 3 diffèrent de ceux du tableau 2 car les expériences faites après la campagne CoNLL 2017 ont permis de mieux utiliser les vingt phrases fournies pour la mise au point du système et surtout de tester différentes combinaisons de langues afin d'identifier le modèle le plus per-

formant pour la tâche. On voit que le système de base est ainsi plus performant que le système officiel ayant participé à CoNLL 2017, même sans ressources extérieures ni modèle multilingue.

	Corpus utilisés	**Score LAS**	**Score UAS**
1	sme (20)	32,96	46,85
2	eng (12 217)	32,72	50,44
3	fin (12 543)	40,74	54,24
4	sme (20) + eng (12 217)	46,54	61,61
5	**sme (20) + fin (12 543)**	**51,54**	**63,06**

Tableau 3. *Évaluation de l'analyse du same du nord (sme) : scores LAS (*labeled attachment scores*) et UAS (*unlabeled attachment scores*), c'est-à-dire scores calculés en prenant en compte l'étiquette de la relation (score LAS, colonne de gauche), et sans elle (score UAS, colonne de droite). La première ligne sme (20) réfère à l'expérience utilisant uniquement sur les vingt phrases annotées de same disponibles pour l'entraînement. Les autres lignes montrent les résultats avec différentes combinaisons de corpus annotés : anglais (eng) et finnois (fin). Pour chaque corpus, le nombre de phrases utilisées est indiqué entre parenthèses.*

Globalement, les expériences que nous avons menées en utilisant le finnois comme source de connaissances (en particulier pour élaborer des plongements de mots bilingues) ont permis d'obtenir de meilleurs résultats qu'avec d'autres langues (on obtient de meilleurs résultats avec le finnois qu'avec l'anglais pour l'analyse syntaxique du same du nord, cf. tableau 3, et on obtient également de meilleurs résultats avec le finnois qu'avec l'anglais pour l'analyse du komi, cf. tableau 4 – le russe est toutefois plus performant pour analyser le komi que le finnois). Ceci semble indiquer d'une

	Corpus utilisés	**Score LAS**	**Score UAS**
1	kpv (10)	22,33	51,78
2	eng (12 217)	44,47	59,29
3	rus (3 850)	53,85	71,29
4	fin (12 543)	48,22	66,98
5	kpv (10) + eng (12 217)	50,47	66,23
6	kpv (10) + rus (3 850)	53,10	69,98
7	kpv (10) + fin (3 850)	53,66	71,29
8	kpv (10) + fin (12 543)	55,16	**73,73**
9	kpv (10) + eng (12 217) + fin (12,543)	52,50	68,57
10	**kpv (10) + rus (3 850) + fin (12 543)**	**56,66**	71,86

Tableau 4. *Évaluation de l'analyse syntaxique du komi. La première ligne kpv (10) réfère à l'expérience utilisant uniquement les dix phrases annotées de komi disponibles pour l'entraînement. Les autres lignes montrent les résultats avec différentes combinaisons de corpus annotés : anglais (eng), russe (rus), et finnois (fin). Pour chaque corpus, le nombre de phrases utilisées est indiqué entre parenthèses.*

part qu'il est possible d'inférer des connaissances linguistiques utiles pour l'analyse à partir d'une langue tierce et, d'autre part, que la typologie et la situation linguistique jouent un rôle (on obtient de meilleurs résultats sur le same ou le komi à partir du finnois ou éventuellement du russe qu'à partir de l'anglais, même si les données utilisées pour l'anglais sont largement plus volumineuses).

La stratégie de transfert de connaissances d'une langue vers l'autre a bien fonctionné pour le finnois vis-à-vis du same, ce qui peut sembler logique car le finnois et le same sont supposés relativement proches génétiquement parlant (mais n'importe quel locuteur pourra aussi dire à quel point ces langues sont éloignées : il n'y a pas d'intercompréhension, même limitée, et même le vocabulaire de base est très différent). Le transfert fonctionne bien aussi pour le komi, alors que le komi est supposé plus éloigné du finnois d'un point de vue génétique.

Concernant le komi, une des hypothèses que nous avions faites *a priori* était qu'un modèle élaboré à partir du russe pourrait aboutir à de meilleurs résultats qu'un modèle acquis à partir du finnois, car le russe a beaucoup « contaminé » le komi depuis plusieurs décennies (du fait de la situation linguistique et du bilinguisme de tous les locuteurs komis). Les résultats pour le russe sont, de fait, étonnants (ligne 3 du tableau 4). On obtient ainsi d'excellents résultats, qui surpassent même les résultats obtenus en ajoutant les dix phrases annotées de komi lors de l'apprentissage (ligne 6). Ceci est en fait sans doute dû à la proximité du russe et du komi, à la présence de cognats et surtout, aux conditions dans lesquelles sont faites ces expériences. L'apprentissage d'analyseurs avec si peu de données est donc possible, mais il faut garder à l'esprit que les résultats sont alors relativement instables (ceci pose d'ailleurs la question de la validité des résultats obtenus à partir d'échantillons aussi petits). Il est probable que, dans d'autres conditions expérimentales, les résultats obtenus avec le corpus russe seul seraient moins bons.

Si l'on fait abstraction des performances obtenues pour le komi à partir du corpus russe seul, nos résultats montrent que l'ajout de phrases annotées de la langue à analyser, même en petite quantité, peut améliorer significativement les résultats obtenus (lignes 5 à 10 ; on l'avait déjà vu lors de notre participation officielle à la tâche commune CoNLL 2017, où l'usage des vingt phrases fournies par défaut, et le fait d'en réserver une partie ou non comme corpus de développement, pouvaient avoir un effet très positif sur les résultats finaux). La taille des corpus bruts utilisés pour l'obtention des plongements de mots, et les ressources annexes utilisées pour l'acquisition de connaissances linguistiques jouent aussi un rôle important (on comparera ainsi les lignes 7 et 8, où le modèle est à chaque fois conçu à partir du finnois, mais avec deux corpus de tailles différentes). C'est logiquement le corpus le plus grand qui permet d'obtenir le modèle le plus performant. On peut aussi comparer les performances relatives obtenues avec des modèles élaborés à partir du russe, du finnois et de l'anglais : l'anglais, même avec un ensemble de phrases annotées très supérieur, n'est pas vraiment compétitif sur le komi. Comme on l'a dit, ces résultats sont toutefois fragiles et sont validés sur un corpus très petit. Il faut donc souhaiter davantage d'études sur des

langues variées, afin d'obtenir une meilleure vue des types de résultats possibles en fonction des langues, des ressources et des algorithmes utilisés.

Finalement, c'est le modèle élaboré à partir du finnois et du russe qui permet d'obtenir les meilleurs résultats (et non celui élaboré à partir de l'anglais, même si on dispose de plus de données pour l'anglais). Il semble bien que les langues choisies pour l'apprentissage jouent un rôle, et il est important de choisir celles-ci en fonction de critères linguistiques et typologiques.

On observe enfin que les scores UAS (c'est-à-dire sans tenir compte des étiquettes de relations syntaxiques) varient légèrement plus que les scores LAS (scores avec les étiquettes des relations syntaxiques), autrement dit les relations de base ont été trouvées et la racine correctement identifiée dans un certain nombre de cas, même quand les étiquettes des relations n'ont pas été attribuées correctement. Il est intéressant de noter que le modèle qui obtient le meilleur score UAS est le couple komi-finnois, même si d'autres combinaisons (avec l'ajout du russe notamment) permettent d'obtenir les meilleurs scores LAS.

7. Conclusion

Dans cet article, nous avons présenté une approche fondée sur l'utilisation de modèles multilingues, afin de fournir une analyse syntaxique pour des langues disposant de peu de ressources, et en particulier ne disposant pas de données annotées (à part quelques phrases utilisées comme amorçage, dix à vingt phrases dans les expériences menées ici et lors de la campagne CoNLL 2017). Nous avons montré que l'approche suivie était efficace dans le cas du same et du komi, même si les performances restent évidemment bien en deçà de ce que l'on peut obtenir avec un corpus d'entraînement de grande taille, comme en témoignent nos résultats pour le same lors de la campagne d'évaluation CoNLL 2018 (avec un corpus d'entraînement au format UD). Il s'agit toutefois, à notre avis, d'un cadre intéressant pour aider à produire des corpus annotés pour des langues peu dotées. Le cas du komi est à cet égard un cas d'étude intéressant, dans la mesure où il s'agit d'une langue avec peu de ressources, mais avec des locuteurs intéressés et demandeurs d'outils d'analyse automatique. Ce cadre pose toutefois un problème d'évaluation, en l'absence de données de référence (*gold standard*).

Nous avons observé que les modèles multilingues permettent généralement d'améliorer les performances par rapport à des modèles monolingues. Les langues génétiquement liées semblent être la meilleure source de connaissances (le finnois est ainsi efficace pour l'analyse du same comme du komi), mais la prise en compte de langues de contact semble aussi pertinente (ainsi le russe pour l'analyse du komi), de même que des langues pour lesquelles on dispose tout simplement de gros corpus (comme l'anglais). Une meilleure compréhension de l'apport réel de chaque langue au processus global serait intéressante pour permettre de définir une stratégie plus générale, et surtout reproductible, concernant le développement et l'utilisation de modèles multilingues pour l'analyse syntaxique.

Remerciements

Les auteurs remercient les trois relecteurs anonymes pour leurs suggestions, qui ont permis de largement améliorer cet article. Les travaux décrits ont été en partie effectués dans le cadre du projet LAKME, financé par l'université Paris Sciences et Lettres (IDEX PSL référence ANR-10-IDEX-0001-02). Cette recherche a aussi bénéficié du soutien d'un projet RGNF-CNRS entre le Lattice et l'université d'État des sciences humaines de Russie.

Annexe : contribution à l'élaboration d'un corpus arboré pour le komi

Afin de permettre l'évaluation de l'analyseur décrit dans cet article, un ensemble de phrases en komi ont été annotées au format UD. Ce corpus comprend actuellement environ trois cents phrases, et devrait en contenir mille prochainement. Une première version de ce corpus a été incluse dans la distribution officielle Universal Dependencies d'avril 2018[9]. Au-delà de la réalisation d'un nouveau corpus arboré, plusieurs points peuvent être soulignés, qui nous semblent relativement typiques du cas des langues de terrain et des langues minoritaires.

Les données disponibles sont de deux types très différents. On a d'un côté des sources écrites, parfois anciennes, écrites dans une langue relativement élaborée et littéraire, parfois très éloignée de la langue quotidienne. De l'autre, on dispose de corpus beaucoup plus modestes (en taille) correspondant à des enquêtes de terrain faites par des linguistes. Ce matériau est précieux, car directement issu de travaux linguistiques, il rend compte de la langue réellement utilisée par les locuteurs au quotidien, mais il pose plusieurs difficultés. Des difficultés matérielles d'abord, dans la mesure où ces données sont souvent encodées dans des formats particuliers, qui ne conviennent pas directement à un traitement automatique ; des difficultés liées à la taille des corpus existants, qui rendent difficile l'utilisation de techniques d'apprentissage artificiel par exemple. Enfin, ces corpus sont représentatifs de l'oral : ils posent donc des problèmes particuliers et les outils mis au point pour l'écrit ne sont pas très performants sur ce type de données.

Notre travail se situe dans le cadre d'un effort en cours en vue de fournir des données annotées pour un certain nombre de langues finno-ougriennes. Des corpus arborés sont déjà disponibles pour le finnois, l'estonien et le hongrois. L'année 2017 a vu l'émergence d'un corpus arboré important pour le same du nord (produit en grande partie automatiquement à partir des outils d'analyse mis au point à l'université de Tromsø et non entièrement vérifié manuellement). Un corpus arboré est actuellement en préparation pour l'erzya (langue mordve), ce qui permettra de couvrir à terme une partie non négligeable des langues finno-ougriennes, même si des efforts seront encore nécessaires pour les autres langues. Une tendance similaire est observée pour ce qui

9. Voir le site officiel de l'initiative Universal Dependencies : `http://universaldependencies.org`.

concerne la réalisation de corpus arborés à partir du résultat d'enquêtes de terrain. À notre connaissance, des projets existent par exemple pour le dargwa (langue du Caucase), le pnar (langue austro-asiatique) et le shipibo-konibo (langue du Pérou).

En pratique...

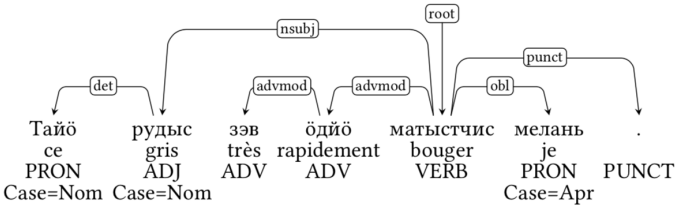
La plupart des travaux sur la langue komie sont actuellement menés à Syktyvkar, Russie (capitale de la République komie). Le FU-Lab, dirigé par Marina Fedina, a en particulier numérisé un nombre important de livres komis (datant du début du XX^{e} siècle jusqu'à aujourd'hui), ceux libres de droits étant directement mis à disposition en ligne. Ce corpus brut compte actuellement quarante millions de mots et l'objectif à long terme est de numériser tous les livres publiés en komi-zyriène. Le nombre total des publications est estimé à environ quatre mille cing cents livres, auxquels il faut ajouter des dizaines de milliers de pages issues de journaux et de revues. Pour la constitution de notre corpus, nous avons évidemment veillé à n'utiliser que des textes libres de droits, afin d'en assurer une distribution aussi large et simple que possible. Il est possible qu'à un stade ultérieur des matériaux moins standard puissent être inclus dans la base de données, par exemple des textes issus de blogs et de discussions en ligne, mais cela pose immédiatement des problèmes de droits et de diffusion.

En dehors du projet mené à Syktyvkar, un des plus grands projets de recherche sur le komi parlé a été un projet de documentation dirigé par Rogier Blokland (université de Uppsala) en 2014-2016. Ce projet a abouti à un grand corpus en langue parlée transcrite. Ces données sont précieuses, pour les raisons que nous avons données *supra*, mais elles sont aussi problématiques car elles ne peuvent généralement pas être diffusées directement dans le domaine public. Le corpus contient aussi des formes dialectales, et comme le komi-zyriène écrit ne suit pas les principes utilisés pour les transcriptions, il semble problématique de mélanger ce matériau avec les données issues de sources écrites. Le corpus oral contient enfin de nombreuses phrases où le komi est mêlé à du vocabulaire russe, les locuteurs pratiquant le code switching en permanence. Cette langue est donc non standard, mais elle est par ailleurs scientifiquement intéressante et pertinente.

À partir de ce point de départ, il a été décidé de créer deux corpus différents, le premier avec les matériaux écrits et le deuxième avec les données orales issues d'enquêtes de terrain.

Annotation syntaxique du corpus komi-zyriène au format UD

Pour l'annotation du corpus komi-zyriène, nous nous sommes inspirés de corpus arborés existants et des consignes d'annotation liées, notamment celles ayant été constituées pour le finnois, le same du nord et l'erzya, ainsi que pour le russe. Il s'agit de langues proches du komi (langues de la même famille, à l'exception du russe), et il nous semblait naturel d'aller voir du côté de ces langues en priorité. De fait, les



ID phrase: belyx-011.042
Traduction: Ce (nuage) gris est venu très rapidement vers moi.

consignes d'annotation ont facilement pu être transposées au komi et quasiment toutes les configurations observées correspondaient à des cas de figure observés (*mutatis mutandis*) dans au moins une de ces langues.

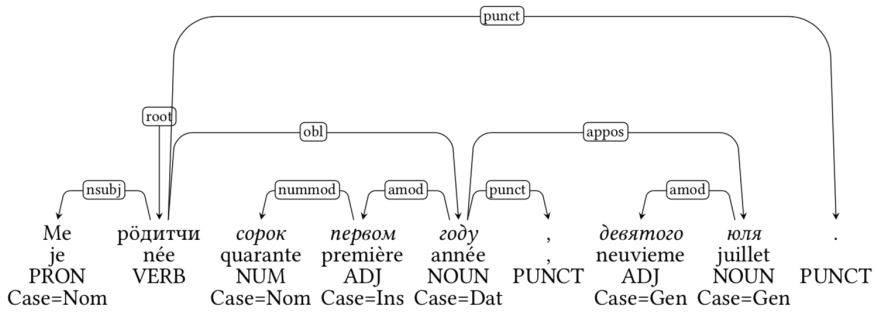
Le komi-zyriène présente malgré tout quelques particularités et différences qui le distinguent de ces langues proches. Il existe notamment deux cas spatiaux largement spécifiques au komi (en fait aux langues permiennes, branche des langues finno-ougriennes qui regroupe le komi et l'oudmourte) : l'égressif et l'approximatif. Ces deux cas expriment le mouvement depuis et vers une direction. Ils se distinguent de l'élatif et de l'illatif (deux cas bien répandus dans l'ensemble des langues finno-ougriennes) : l'élatif et l'illatif expriment aussi un mouvement depuis ou vers un lieu, mais ils insistent justement sur ce point de départ ou d'arrivée, alors que l'égressif et l'approximatif insistent davantage sur le mouvement, sans préciser le point de départ ou d'arrivée.

L'exemple ci-dessous illustre précisément l'utilisation du cas approximatif, sans aucun caractère égressif.

Il existe deux autres cas directionnels en komi, traditionnellement appelés prolatifs et transitifs, qui expriment le mouvement le long d'un chemin. Ceux-ci correspondent assez bien au cas appelé « perlatif » du guide d'annotation au format UD (Universal Dependencies), mais ce cas ne figure pas dans les exemples annotés jusqu'ici. Plus généralement, ceci pose la question de la terminologie employée et de la mise en rapport des cas (et plus généralement des notions linguistiques) à travers les langues : nous pensons que le prolatif et le translatif correspondent au perlatif, mais ceci mériterait sûrement une discussion approfondie. UD est en tout cas l'occasion de s'interroger sur la terminologie en cours, les notions manipulées et les correspondances entre langues. Il était à cet égard utile de garder un œil sur le russe lors de l'annotation, dans la mesure où le komi emprunte certaines constructions au russe. De plus, il semble souhaitable d'être, autant que faire se peut, cohérent et homogène au niveau des annotations.

Le premier exemple illustre une situation typique où la date est exprimée en russe, y compris au niveau du marquage morphologique et syntaxique.

Nous avons cherché ici à avoir une annotation comparable à celle de la structure correspondante en russe. Ce choix diffère de ce qui a été fait pour la plupart des autres



ID phrase: kpv_izva20150403-2-b.014
Traduction: Je suis née (l'année) 1941, le 9 juillet.

langues, où les mots ou structures en langue étrangère sont généralement marqués en tant que tels, et donc globalement plutôt mis de côté. Vu le bilinguisme de tous les locuteurs komis qui se retrouve en partie dans les corpus, nous souhaitons avoir une annotation qui intègre pleinement les passages en russe (y compris à l'intérieur d'une même phrase en cas de code switching comme ici) et les considère comme faisant pleinement partie du corpus komi.

Il faut toutefois noter que ceci peut aussi entraîner différents problèmes. Par exemple, certaines structures (provenant du russe) auront un trait `gender` (exprimant le genre grammatical), alors qu'il s'agit d'un trait morphologique étranger au komi. Ceci est évidemment aussi un défi pour les outils de TAL et les analyseurs en général, qui doivent gérer des situations linguistiques plus complexes que ce que l'on trouve dans la plupart des grands corpus monolingues disponibles. C'est cette richesse qui fait l'intérêt de ces langues trop longtemps laissées de côté.

8. Bibliographie

Aikio A., « An essay on Saami ethnolinguistic prehistory », *in* R. Grünthal, P. Kallio (eds), *A Linguistic Map of Prehistoric Northern Europe*, Société Finno-Ougrienne, Helsinki, p. 63-117, 2012.

Ammar W., Mulcaire G., Ballesteros M., Dyer C., Smith N. A., « Many Languages, One Parser », *Transactions of the Assoc. for Comp. Linguistics (TACL)*, vol. 4, p. 431-444, 2016a.

Ammar W., Mulcaire G., Tsvetkov Y., Lample G., Dyer C., Smith N. A., « Massively multilingual word embeddings », *Prépublication arXiv :1602.01925*, 2016b.

Artetxe M., Labaka G., Agirre E., « Learning principled bilingual mappings of word embeddings while preserving monolingual invariance », *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, p. 2289-2294, 2016.

Artetxe M., Labaka G., Agirre E., « Learning bilingual word embeddings with (almost) no bilingual data », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Vancouver, p. 451-462Proceedin, 2017.

Ballesteros M., Goldberg Y., Dyer C., Smith N. A., « Training with Exploration Improves a Greedy Stack LSTM Parser », *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, p. 2005-2010, 2016.

Björkelund A., Falenska A., Yu X., Kuhn J., « IMS at the CoNLL 2017 UD Shared Task : CRFs and Perceptrons Meet Neural Networks », *CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, p. 40-51, 2017.

Blokland R., Fedina M., Gerstenberger C., Partanen N., Rießler M., Wilbur J., « Language documentation meets language technology », *in* T. Pirinen, F. Tyers, T. Trosterud (eds), *Workshop on Comp. Linguistics for Uralic Languages*, Tromsø, 2015.

Bojanowski P., Grave E., Joulin A., Mikolov T., « Enriching Word Vectors with Subword Information », *Transactions of the Assoc. for Comp. Linguistics (TACL)*, vol. 5, p. 135-146, 2017.

Che W., Guo J., Wang Y., Zheng B., Zhao H., Liu Y., Teng D., Liu T., « The HIT-SCIR System for End-to-End Parsing of Universal Dependencies », *CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, p. 52-62, 2017.

Chen D., Manning C. D., « A Fast and Accurate Dependency Parser using Neural Networks. », *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, p. 740-750, 2014.

Cho K., « Natural language understanding with distributed representation », *Prépublication arXiv :1511.07916*, 2015.

Conneau A., Lample G., Ranzato M., Denoyer L., Jégou H., « Word Translation Without Parallel Data », *Conf. International Conference on Learning Representations (ICLR)*, Toulon, 2017.

Das D., Petrov S., « Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections », *Conf. Assoc. for Comp. Linguistics (ACL)*, Portland, 2011.

Dozat T., Qi P., Manning C. D., « Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task », *CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, p. 20-30, 2017.

Gerstenberger C., Partanen N., Rießler M., Wilbur J., « Utilizing language technology in the documentation of endangered Uralic languages », *Northern European Journal of Language Technology*, vol. 4, p. 29-47, 2016.

Gouws S., Bengio Y., Corrado G., « BilBOWA : Fast Bilingual Distributed Representations without Word Alignments », *Intern. Conf. on Machine Learning (ICML)*, Lille, p. 748-756, 2015.

Gouws S., Søgaard A., « Simple task-specific bilingual word embeddings », *Conf. of the North American Chapter of the Assoc. for Comp. Linguistics – Human Language Technologies (NAACL-HLT)*, Denver, p. 1386-1390, 2015.

Guo J., Che W., Yarowsky D., Wang H., Liu T., « Cross-lingual Dependency Parsing Based on Distributed Representations », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Beijing, 2015.

Guo J., Che W., Yarowsky D., Wang H., Liu T., « A Representation Learning Framework for Multi-Source Transfer Parsing. », *Conf. of the Association for the Advancement of Artificial Intelligence (AAAI)*, Beijing, p. 2734-2740, 2016.

Huang Z., Xu W., Yu K., « Bidirectional LSTM-CRF Models for Sequence Tagging », *Prépublication arxiv1508.01991*, 2015.

Hwa R., Resnik P., Weinberg A., Cabezas C., Kolak O., « Bootstrapping Parsers via Syntactic Projection Across Parallel Texts », *Natural Language Engineering*, vol. 11, n° 3, p. 311-325, 2005.

Khapra M. M., Sohoney S., Kulkarni A., Bhattacharyya P., « Value for Money : Balancing Annotation Effort, Lexicon Building and Accuracy for Multilingual WSD », *Coling*, Beijing, p. 555-563, 2010.

Kim S., Jeong M., Lee J., Lee G. G., « A Cross-lingual Annotation Projection Approach for Relation Detection », *Coling*, Beijing, p. 564-571, 2010.

Kim S., Jeong M., Lee J., Lee G. G., « Cross-Lingual Annotation Projection for Weakly-Supervised Relation Extraction », *ACM Transactions on Asian Language Information Proc. (TALIP)*, vol. 13, n° 1, p. 3 :1-3 :26, February, 2014.

Kiperwasser E., Goldberg Y., « Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations », *Transactions of the Assoc. for Comp. Linguistics (TACL)*, vol. 4, p. 313-327, 2016.

Kozhevnikov M., Titov I., « Cross-lingual bootstrapping for semantic role labeling », *xLiTe : Cross-Lingual Technologies*, Lake Tahoe, 2012.

Leinonen M., « The Russification of Komi », *The Slavicization of the Russian North. Mechanisms and Chronology*, Slavica Helsingiensia 27, p. 234-245, 2006.

Levy O., Søgaard A., Goldberg Y., « A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments », *Conf. of the European Chapter of the Assoc. for Comp. Linguistics (EACL)*, Valence, p. 765-774, 2017.

Lim K., Partanen N., Poibeau T., « Multilingual Dependency Parsing for Low-Resource Languages : Case Studies on North Saami and Komi-Zyrian », *Language Resource and Evaluation Conference (LREC)*, Miyazaki, 2018.

Lim K., Poibeau T., « A System for Multilingual Dependency Parsing based on Bidirectional LSTM Feature Representations », *CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, p. 63-70, 2017.

Liu K., Lü Y., Jiang W., Liu Q., « Bilingually-Guided Monolingual Dependency Grammar Induction », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Sofia, p. 1063-1072, 2013.

Luong T., Pham H., Manning C. D., « Bilingual Word Representations with Monolingual Quality in Mind. », *Conf. of the North American Chapter of the Assoc. for Comp. Linguistics – Human Language Technologies (NAACL-HLT)*, Denver, p. 151-159, 2015.

McDonald R., Crammer K., Pereira F., « Online large-margin training of dependency parsers », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Ann Arbor, p. 91-98, 2005a.

McDonald R., Pereira F., Ribarov K., Hajič J., « Non-projective Dependency Parsing Using Spanning Tree Algorithms », *Conf. on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, Stroudsburg, p. 523-530, 2005b.

Mikolov T., Chen K., Corrado G., Dean J., « Efficient Estimation of Word Representations in Vector Space », *Prépublication arXiv :1301.3781*, 2013a.

Mikolov T., Le Q. V., Sutskever I., « Exploiting similarities among languages for machine translation », *Prépublication arXiv :1309.4168*, 2013b.

Naseem T., Barzilay R., Globerson A., « Selective sharing for multilingual dependency parsing », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Bruxelles, p. 629-637, 2012.

Nivre J., « Incrementality in deterministic dependency parsing », *Workshop on Incremental Parsing (organisé avec la conf. ACL 2004)*, Barcelone, p. 50-57, 2004.

Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajič J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D., « UDPipe : Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing », *Language Resources and Evaluation Conf. (LREC)*, Portorož, 2016.

Ruder S., Vulić I., Søgaard A., « A survey of cross-lingual embedding models », *Prépublication arXiv :1706.04902*, 2017.

Scherrer Y., Sagot B., « A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages », *Language Resources and Evaluation Conf. (LREC)*, Reykjavik, 2014.

Shi T., Wu F. G., Chen X., Cheng Y., « Combining Global Models for Parsing Universal Dependencies », *CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, p. 31-39, 2017.

Smith D. A., Eisner J., « Parser Adaptation and Projection with Quasi-synchronous Grammar Features », *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Singapour, p. 822-831, 2009.

Straka M., Hajič J., Straková J., « UDPipe : Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing », *Language Resources and Evaluation Conf. (LREC)*, Portorož, 2016.

Täckström O., McDonald R., Uszkoreit J., « Cross-lingual word clusters for direct transfer of linguistic structure », *Conf. of the North American chapter of the Assoc. for Comp. Linguistics (NAACL)*, Montréal, p. 477-487, 2012.

Taskar B., Chatalbashev V., Koller D., Guestrin C., « Learning structured prediction models : A large margin approach », *Int. Conf. on Machine Learning (ICML)*, Bonn, p. 896-903, 2005.

Vulić I., Søgaard A., Ruder S., « On the Limitations of Unsupervised Bilingual Dictionary Induction », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Melbourne, 2018.

Weiss D., Alberti C., Collins M., Petrov S., « Structured Training for Neural Network Transition-Based Parsing », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Beijing, 2015.

Zeman D., Hajič J., Popel M., Potthast M., Straka M., Ginter F., Nivre J., Petrov S., « CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies », *CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Bruxelles, p. 1-20, 2018.

Zeman, D. *et al.*, « CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies », *CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, p. 1-19, 2017.

Zhao H., Song Y., Kit C., Zhou G., « Cross Language Dependency Parsing Using a Bilingual Lexicon », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, p. 55-63, 2009.

Zhuang T., Zong C., « Joint Inference for Bilingual Semantic Role Labeling », *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Cambridge, USA, p. 304-314, 2010.