Transformation d'annotations en parties du discours et lemmes vers le format Universal Dependencies : étude de cas pour l'alsacien et l'occitan

Aleksandra Miletic, Delphine Bernhard, Myriam Bras, Anne-Laure Ligozat, Marianne Vergez-Couret

Transformation d'annotations en parties du discours et lemmes vers le format Universal Dependencies : étude de cas pour l'alsacien et l'occitan

Aleksandra Miletić[1] Delphine Bernhard[2] Myriam Bras[1]
Anne-Laure Ligozat[3] Marianne Vergez-Couret[4]
(1) CLLE-ERSS, CNRS, Université Toulouse-Jean Jaurès
(2) LiLPa - EA 1339, Université de Strasbourg
(3) LIMSI, CNRS, ENSIIE, Université Paris-Saclay, F-91405 Orsay
(4) FoReLLIS - EA 3816, Université de Poitiers
aleksandra.miletic@univ-tlse2.fr, dbernhard@unistra.fr,
myriam.bras@univ-tlse2.fr, annlor@limsi.fr,
marianne.vergez.couret@univ-poitiers.fr

RÉSUMÉ

Cet article présente un retour d'expérience sur la transformation de corpus annotés pour l'alsacien et l'occitan vers le format CONLL-U défini dans le projet *Universal Dependencies*. Il met en particulier l'accent sur divers points de vigilance à prendre en compte, concernant la tokénisation et la définition des catégories pour l'annotation.

ABSTRACT

Converting POS-tag and Lemma Annotations into the Universal Dependencies Format : A Case Study on Alsatian and Occitan

This article presents a retrospective report on the transformation of annotated corpora for Alsatian and Occitan into the CONLL-U format defined in the Universal Dependencies project. In particular, it emphasizes various issues to be taken into account, concerning the tokenization and the definition of the categories.

MOTS-CLÉS : annotation, alsacien, occitan, Universal Dependencies.

KEYWORDS: annotation, Alsatian, Occitan, Universal Dependencies.

1 Introduction

Les langues régionales de France sont à l'heure actuelle encore largement sous-dotées en ressources linguistiques, qu'il s'agisse de corpus, bruts ou annotés, ou de ressources lexicales comme des lexiques flexionnels ou des dictionnaires bilingues. Nous nous intéressons dans cet article à l'alsacien et à l'occitan qui, bien que se trouvant dans des situations différentes (famille linguistique, vitalité, ressources existantes), font face à des défis communs. En effet, le développement de ressources et d'outils pour les langues peu dotées nécessite une approche pragmatique qui prend en compte les faibles moyens financiers et humains généralement disponibles. Ainsi, Soria *et al.* (2013) ont énoncé un ensemble de principes mettant l'accent sur la coopération, l'utilisation de standards internationaux, la réutilisation d'approches et de ressources existantes, le partage des ressources et outils produits dans des formats ouverts.

Dans cet article nous montrons comment nous avons repris ces divers principes à notre compte pour le

développement de corpus annotés en parties du discours (POS) et lemmes dans le format CONLL-U [1] défini dans le projet *Universal Dependencies* (UD) (Nivre *et al.*, 2016), pour l'alsacien et l'occitan. Le choix du format CONLL-U répond à au moins deux des principes énoncés par Soria *et al.* (2013) : utilisation de standards et réutilisation d'approches existantes. Par ailleurs, les dialectes alsaciens et l'occitan ne sont pour l'heure pas représentés dans les corpus *Universal Dependencies*. Nous souhaitons pouvoir combler ce manque en partageant les ressources annotées produites avec une licence libre. Cet objectif respecte deux autres principes de Soria *et al.* (2013), à savoir le partage des ressources sur une plateforme pérenne et avec une licence libre, et permet en outre de donner une meilleure visibilité à ces langues. Par ailleurs, le format CONLL-U peut être utilisé directement pour entraîner divers outils (par exemple, spaCy [2] ou UDPipe (Straka & Straková, 2017)), ce qui constitue également un atout pour les langues disposant de peu de moyens pour développer de nouveaux outils.

Nous mettons aussi l'accent sur divers points de vigilance qui sont à prendre en compte lors du passage de corpus annotés vers le format CONNL-U et le jeu d'étiquettes morphosyntaxiques définies dans le projet *Universal Dependencies*. Nous souhaitons ainsi que ce retour d'expérience puisse servir pour d'autres langues, et notamment des langues peu dotées. Même si ce travail a déjà été réalisé pour d'autres langues mieux dotées, cela ne le rend pas trivial pour autant, et ce d'autant plus que les comptes rendus de ce type d'expériences ont peu été publiés.

2 Annotation de corpus au format UD

Le projet *Universal Dependencies* vise à proposer des corpus annotés de manière cohérente pour différentes langues, grâce à des principes d'annotation communs et des catégories unifiées pour les parties du discours, les propriétés lexicales et grammaticales des mots, et les relations de dépendance syntaxique (Nivre *et al.*, 2016). Depuis les origines du projet, le nombre de corpus arborés et de langues répertoriées ne cesse de croître. On trouve notamment des langues considérées comme peu dotées, représentées par des corpus de petite taille (quelques milliers de tokens) comme le breton, le féroïen ou encore le komi-zyriène.

Certains corpus sont directement annotés en suivant les catégories pré-définies dans *Universal Dependencies*, comme par exemple le corpus komi-zyriène produit par le Lattice pour des besoins d'évaluation (Lim *et al.*, 2018). Cependant, d'autres standards existent, comme par exemple le standard GRACE (Rajman *et al.*, 1997), lui-même dérivé des jeux d'étiquettes MULTEXT (Ide & Véronis, 1994) et EAGLES (von Rekowski, 1996). Ils ont été très largement employés pour l'annotation de multiples corpus dans plusieurs langues, dont l'occitan, l'une des langues étudiées dans cet article (cf. section 3).

De nombreux corpus UD ont ainsi été produits par transformation semi-automatique à partir de corpus étiquetés et arborés existants, à l'aide notamment de tables de conversion entre jeux d'étiquettes [3]. Cela étant, ce processus de conversion peut être plus complexe et nécessiter des opérations supplémentaires. Nous nous intéressons ici plus particulièrement aux étapes de découpage en tokens et à l'annotation en lemmes et parties du discours qui concernent directement les travaux présentés dans cet article. Concernant ces aspects, Chun *et al.* (2018) décrivent le processus pour trois corpus arborés du coréen. La transformation vers le format UD version 2.0 a nécessité un redécoupage en tokens de l'un des corpus, pour les tokens incluant des signes de ponctuation et des symboles. Pour les deux autres corpus, des tables de correspondance ont été établies entre les étiquettes morphosyntaxiques

1. `http://universaldependencies.org/docs/format.html`
2. `https://spacy.io/`
3. Voir `https://universaldependencies.org/tagset-conversion/index.html`

existantes et celles du projet UD. Sanguinetti *et al.* (2018) décrivent le processus de conversion vers UD d'un corpus de tweets en italien. Certains tokens ont dû être redécoupés, à savoir les contractions préposition-article et verbe-clitique et un certain nombre de lemmes ont été ajoutés manuellement, car ne pouvant faire l'objet d'une lemmatisation automatique. Les formes non standards (mots étrangers ou dialectaux, formes tronquées ou amalgamées) ont été étiquetées avec la catégorie `X`.

Les objectifs d'annotation unifiée à travers les langues du projet UD posent bien évidemment divers problèmes qui ont été soulignés à plusieurs reprises. L'annotation en relations de dépendance syntaxique impose un découpage particulier en tokens (voir (Gerdes & Kahane, 2016) pour une discussion de ces choix). Certains mots *orthographiques*, tels que produits généralement par les outils de tokénisation, doivent être découpés en plusieurs unités. C'est notamment le cas pour les formes contractées ou des formes non standard, que l'on peut trouver dans les contenus produits par les utilisateurs sur le web (Alonso *et al.*, 2016). Ce découpage particulier en tokens a d'ailleurs été un des problèmes auxquels nous avons été confrontées (voir section 4). Par contre, l'annotation en lemmes et parties du discours est moins problématique en raison de la flexibilité du format. Nous avons tout de même choisi de ne pas respecter systématiquement les définitions proposées pour chaque partie du discours en raison des spécificités des langues traitées (voir section 5.3).

3 Jeux d'étiquettes originaux pour l'alsacien et l'occitan

Les jeux d'étiquettes originaux relèvent de deux situations différentes : celui utilisé pour les dialectes alsaciens était déjà très proche des *Universal POS tags* car s'en inspirant fortement, tandis que celui de l'occitan s'inspirait des catégories plus nombreuses de GRACE (Rajman *et al.*, 1997). En plus des parties du discours, d'autres informations ont également été ajoutées lors de l'annotation : le lemme et sa traduction en français, ainsi que les noms de lieux. Ces annotations supplémentaires permettent de constituer directement des lexiques bilingues à partir des corpus, et d'évaluer l'annotation en entités de type lieu. Les corpus annotés sont de taille sensiblement similaire pour les deux langues : environ 12 600 tokens pour l'alsacien et 11 900 pour l'occitan (Bernhard *et al.*, 2018c).

Les catégories initiales pour l'alsacien sont très proches des catégories proposées dans UD, respectant ainsi le principe de réutilisation soutenu par Soria *et al.* (2013). Il existe peu de descriptions linguistiques détaillées de la morphosyntaxe de l'alsacien et il nous semblait donc plus réaliste de partir d'un nombre limité de catégories bien décrites, compte tenu également du temps limité et des moyens à disposition. Nous avons toutefois ajouté les catégories suivantes aux 17 catégories UD (Bernhard *et al.*, 2018b) : `EPE` pour les épenthèses (insertions de sons pour faciliter l'articulation), `APPRART` pour les contractions préposition + article, `MOD` pour les modaux et `FM` pour les mots d'une autre langue (souvent le français). Le cas des épenthèses est discuté dans la section 5.2 et celui des contractions le sera dans la section suivante. Les catégories `MOD` et `FM` ont été ajoutées car relativement claires et simples à annoter à ce stade.

Pour l'occitan, deux ressources ont été produites : Loflòc, lexique de formes fléchies (Vergez-Couret, 2016; Bras *et al.*, 2017) puis un corpus annoté en catégories morphosyntaxiques, avec un même jeu d'étiquettes, adapté du standard GRACE (Bras *et al.*, 2018). Les étiquettes GRACE peuvent être interprétées comme des étiquettes à 3 niveaux. Le premier niveau d'étiquette permet d'indiquer la catégorie grammaticale des formes fléchies, de classer les signes de ponctuation (`F`) et les formes attestées dont la classification n'a pas encore été réalisée (`X`). Le deuxième niveau propose une classification sémantique ou fonctionnelle spécifique à chaque catégorie de niveau 1. Le troisième niveau concerne les informations morphosyntaxiques flexionnelles, telles le genre, le nombre, la personne, le temps verbal, etc. Des modifications ont été apportées par rapport au jeu d'étiquettes

GRACE. Par exemple, l'ajout, pour les verbes, d'un attribut "Form" pouvant prendre les valeurs positif/négatif afin d'annoter l'impératif négatif de l'occitan dont la forme se distingue de celle de l'impératif positif. Une description détaillée du jeu d'étiquettes est disponible dans le guide d'annotation (Bras, 2018).

tokens	Cossí	aquò	pòt	èsser	?
VO	Rx	Pd	Vm	Vm	F
UD	ADV	PRON	VERB	VERB	PUNCT
lemme	cossí	aquò	poder	èsser	?
glose	comment	ça	pouvoir	être	?

FIGURE 1 – Phrase annotée en occitan. VO = étiquettes originales, UD = étiquettes après conversion.

Dans la version actuelle du corpus, nous retenons les deux premiers niveaux d'étiquettes (le POS et la sous-catégorie sémantique ou fonctionnelle, cf. Figure 1), ce qui donne un jeu de 40 étiquettes au total ; ces annotations ont été corrigées manuellement après une première annotation automatique. Le corpus a également été étiqueté en informations flexionnelles, mais cette couche d'information, produite de manière automatique, n'a pas encore été validée par des annotateurs humains. Ceci fait partie des pistes pour la suite du travail. Le choix de ce jeu d'étiquettes est principalement motivé par le fait d'exploiter des ressources existantes pour l'occitan (le lexique mentionné ci-dessus), mais aussi pour d'autres langues proches, notamment pour le français (cf. FTB (Abeillé *et al.*, 2003), construit en utilisant les mêmes standards) et pour le catalan (cf. AnCora-CA (Taulé *et al.*, 2008), fondé sur des principes similaires).

Pour rendre nos corpus compatibles avec les exigences du projet UD, il a été nécessaire d'intervenir à deux niveaux : il a fallu adapter le découpage en tokens (cf. section 4) et ensuite convertir l'annotation morphosyntaxique vers le jeu d'étiquettes UD (cf. section 5). Le travail s'est ici partagé entre les deux équipes spécialistes de l'alsacien et de l'occitan, pour les questions plus linguistiques, et la réalisation pratique a été supervisée par le LIMSI concernant la mise en oeuvre des scripts de conversion et de vérification. Ce partage du travail a permis une mise en commun des solutions aux problèmes rencontrés, garantissant ainsi une meilleure qualité aux ressources produites.

4 Transformation vers UD : segmentation en tokens

Comme il a été mentionné dans la section 2, le guide de tokénisation UD pose des exigences spécifiques [4]. Tout d'abord, il considère comme unité textuelle de base un token *syntaxique* et non pas *orthographique*. Cela implique, par exemple, la séparation des formes contractées *préposition + article*, présentes aussi bien en alsacien qu'en occitan. En même temps, le projet proscrit les mots « multi-tokens » (incluant des espaces) et les expressions polylexicales (*multiword expressions*) sont systématiquement traitées par l'annotation syntaxique plutôt que par la tokénisation, sauf dans quelques cas exceptionnels.

4.1 Formes contractées

Le découpage d'un texte en tokens syntaxiques n'est pas une question anodine. Gerdes & Kahane (2016) notent que cette approche favorise le principe d'adéquation de l'annotation linguistique, mais qu'elle est contraire au principe de simplicité, vu qu'elle peut entraîner le besoin d'une validation manuelle de la tokénisation. Et dans le cadre d'une conversion d'un corpus existant, il faut non seulement opérer les découpages, mais aussi fournir une annotation pour les formes obtenues.

4. Voir `https://universaldependencies.org/u/overview/tokenization.html`

Même dans le cas d'une langue avec un nombre limité de ces formes, ce passage peut se montrer problématique. En décrivant la conversion du corpus catalan, Alonso & Zeman (2016) indiquent que le catalan dispose de 6 formes contractées (*al*, *als*, *del*, *dels*, *pel*, *pels*) et décrivent leur découpage et leur ré-annotation. Or, dans le corpus distribué (v2.3), on retrouve un nombre élevé de ces formes non traitées (au-delà de 14 000 dans l'ensemble du corpus). En revanche, la séparation des clitiques semble bien effectuée (cf. *ofrir-los* traité comme deux formes séparées, *ofrir* et *los*). L'occitan languedocien dispose des mêmes formes contractées que le catalan, plus les formes *sul*, *suls*, *jol*, *jols*, *vèl* et *vèls*. Pour les autres dialectes, les formes contractées diffèrent légèrement mais il est également possible de les lister. Elles portent des étiquettes spéciales concaténées (`SpDa`), ce qui facilite leur repérage et l'identification des étiquettes à attribuer aux formes découpées (*al* `SpDa` → *a* `Sp` + *lo* `Da`).

Pour l'alsacien en revanche, il existe une grande variabilité dans les graphies de ces formes : on trouve par exemple les variantes *zuem*, *züem*, et *zum* pour *zu* + *dem*. Il est donc difficile d'en établir une liste *a priori* et, par conséquent, de les identifier lors de la tokénisation d'un texte brut. Comme pour l'occitan, nous disposons d'une étiquette POS dédiée, ce qui facilitait le repérage des formes contractées. En effet, le jeu d'étiquettes initial intègre la catégorie `APPRART` pour les contractions préposition + article. Cette catégorie est directement tirée du jeu d'étiquettes STTS pour l'allemand (Schiller *et al.*, 1999). Le guide d'annotation STTS nous a été très utile pour constituer notre propre guide et résoudre des difficultés rencontrées lors de l'annotation manuelle. Nous avons procédé à un découpage semi-automatique des tokens annotés `APPRART` dans le corpus alsacien, afin de les segmenter en deux tokens, l'un étiqueté `ADP`, l'autre `DET`. En raison de la variation graphique, 40 formes de ce type ont été repérées dans le corpus. La Figure 2 donne l'exemple d'une phrase annotée en alsacien. Dans cet exemple, *Mìtem* a été segmenté en *Mìt* + *dem*.

tokens	Mìtem		Sabayon	ìwwerzìehje	ùn	mìt	de	g'hobelte	Màndle	bstraie	.
VO	APPRART		NOUN	VERB	CONJ	ADP	DET	ADJ	NOUN	VERB	PUNCT
UD	ADP	DET	NOUN	VERB	CCONJ	ADP	DET	ADJ	NOUN	VERB	PUNCT
lemme	mìt	de	Sabayon	ìwwerzìehje	ùn	mìt	de	g'hobelt	Màndel	bstraie	.
glose	avec	le	sabayon	napper	et	avec	les	effilé	amande	saupoudrer	.

FIGURE 2 – Phrase annotée en alsacien. VO - étiquettes originales, UD - étiquettes après conversion.

4.2 Traitement des mots « multi-tokens »

L'exigence d'UD de ne pas utiliser de mots « multi-tokens » ne concerne que le corpus occitan. À la différence du corpus alsacien, découpé en tokens orthographiques, le corpus occitan contient un certain nombre d'unités polylexicales qui ont été soudées en un seul token lors de l'annotation manuelle. Il s'agit notamment de séquences figées qui n'ont pas de lecture libre, le plus souvent des locutions adverbiales ou conjonctives qui peuvent s'écrire aussi en un seul mot graphique (*ça_que_la* 'pourtant', *si_que_non* 'sinon'). La liste complète de ces formes est disponible dans le guide d'annotation (Bras, 2018). Le repérage et le découpage automatique de ces formes n'est pas problématique, vu l'utilisation systématique du caractère "_". En revanche, l'attribution des étiquettes aux formes découpées peut l'être : certains de ces figements ne sont pas transparents au niveau de leur structure syntaxique et l'identification des parties du discours de leurs éléments n'est pas intuitive, même pour un annotateur humain (cf. la forme *ça_que_la*, annotée PRON SCONJ ADV, où la reconnaissance de *la* comme adverbe de lieu passe par le repérage, pour l'annotateur, de la forme plus fréquente *lai*). Dans la version actuelle du corpus, ces unités se présentent encore comme tokens multi-mots. Elles seront traitées dans l'étape suivante du travail, avant d'aborder l'annotation syntaxique.

5 Transformation vers UD : étiquettes de parties du discours

Les corpus n'ayant pas été annotés dans le format UD dès le départ, il a été nécessaire de convertir les étiquettes initiales. Cette tâche était moins exigeante pour l'alsacien, dont le jeu d'étiquettes original est majoritairement fondé sur celui de UD (cf. section 3). La transformation du jeu d'étiquettes occitan a posé plus de défis, notamment dus à des différences de granularité et de découpage des catégories.

5.1 Définition d'une table de correspondances

Pour la conversion, nous avons utilisé des tables de correspondance dans des scripts de conversion automatique. La difficulté principale a été de définir ces correspondances en prenant en compte les caractéristiques des langues traitées. Le passage vers le jeu d'étiquettes UD nous a permis de vérifier la cohérence des annotations des corpus initiaux (par exemple, cohérence des lemmes pour les nombres en occitan). La détection de ces incohérences a été effectuée de manière automatique et la correction a ensuite été réalisée manuellement.

Pour l'alsacien, nous avons conservé les étiquettes initiales, pour celles se trouvant dans le jeu UD. La catégorie `MOD` devient `AUX`, `FM` devient `X`. Les cas de `APPRART` et `EPE` sont discutés respectivement dans les sections 4.1 et 5.2.

Pour l'occitan, certaines étiquettes avaient des correspondants du même niveau de granularité dans le jeu UD (nom commun : `Nc` → `NOUN`; nom propre : `Np` → `PROPN`; verbe principal : `Vm` → `VERB`; verbe auxiliaire : `Va` → `AUX`). Mais ce passage d'un jeu de 40 étiquettes vers un jeu de 17 étiquettes a également entraîné une perte d'information dans l'étiquette UD (mais pas dans le corpus, qui conserve également l'étiquette d'origine). Différents types d'adverbes (`Rg` - adv. général, `Rx` - adv. interrogatif/exclamatif, `Rq` - adv. d'intensité/de quantité) sont tous exprimés par une seule étiquette - `ADV`. Sur les 8 étiquettes pronominales de départ, 7 sont traduites par `PRON`. La conversion a également soulevé des questions de découpage des catégories : dans le jeu original, les cardinaux sont annotés en fonction de leur comportement syntaxique comme noms, adjectifs, pronoms ou déterminants cardinaux. Or, dans le cadre de UD, les numéraux sont traités comme une seule catégorie – `NUM`.

Néanmoins, les pertes d'information décrites ci-dessus peuvent être neutralisées si l'on adopte également l'annotation en traits lexicaux et flexionnels. Ceci permettrait de rétablir les distinctions entre les étiquettes pronominales : les pronoms personnels, démonstratifs, indéfinis, relatifs et interrogatifs porteraient différentes valeurs du trait `PronType` (`PronType=Prs|Dem|Ind|Rel|Int`), alors que les pronoms possessif et réflexif seraient marqués comme `PronType=Prs`, mais ils porteraient également des traits supplémentaires, respectivement `Poss=Yes` et `Reflex=Yes`.

5.2 Cas de l'épenthèse

Pour le corpus alsacien, nous avons choisi d'annoter les épenthèses avec une étiquette EPE : *fànga␣**-n-/EPE**␣à drucka*. Ce phénomène a été annoté de diverses manières dans d'autres corpus existants : dans la version UD du corpus français Sequoia (Candito & Seddah, 2012), le "-t-" de liaison (dans "semble-t-il" par exemple) est annoté comme PART ; le guide d'annotation du corpus TCOF-POS (Benzitoun *et al.*, 2012) prévoit une étiquette EPE mais elle n'est utilisée qu'une fois dans le corpus ("que **l'/EPE** on ne voit"). Dans notre cas particulier, EPE a été transformé en PART.

5.3 Différences dans les définitions

D'une manière générale, l'existence de catégories qui semblent identiques dans la liste des catégories initiales et la liste UD ne signifie pas nécessairement que ces catégories ont la même définition. Par

exemple, en alsacien, nous avons annoté les particules verbales séparées PART, ce qui ne correspond pas aux principes UD qui recommandent d'étiqueter les particules séparables ADP ou ADV en fonction de leur type d'origine [5]. Nous avons dans ce cas choisi de nous référer aux recommandation STTS pour la catégorie PTKVZ (*abgetrennter Verbzusatz*, classé dans la catégorie *Partikel* - particule). Dans ce cas bien précis, la transformation vers UD voudrait que soit vérifié chaque token étiqueté PART de manière à choisir soit ADP, soit ADV. Ce travail n'a pour l'heure pas été effectué.

Pour l'occitan, l'expression de la possession dans le groupe nominal pouvant être réalisée par le déterminant *mon* (cf. *mon filh*) ou l'adjectif *miu* (cf. *lo mieu filh* ou *lo filh mieu*), l'étiquette As (adj. possessif) a été traduite par ADJ et non par DET, contrairement aux préconisations du guide UD.

5.4 Informations additionnelles

Enfin, les informations additionnelles produites lors de l'annotation sont ajoutées dans la dernière colonne du fichier CONLL-U (glose en français et informations sur les entités de lieux). Comme prévu dans le format CONLL-U, nous avons également conservé l'annotation d'origine dans une colonne.

6 Conclusion et perspectives

Nous avons présenté un retour d'expérience sur la transformation de deux corpus annotés en alsacien et occitan vers le format UD. Notre travail de réflexion sur l'annotation a démarré au moment de la publication de UD v1 (Nivre *et al.*, 2016). Entre temps, le projet UD a pris de l'ampleur, et une version 2 des recommandations a été publiée. Il nous a donc semblé opportun de transformer nos corpus annotés vers le format préconisé par UD, pour des questions de visibilité et de partage de nos ressources, ainsi que pour aborder l'étape suivante de l'analyse syntaxique en s'appuyant sur les travaux réalisés dans des langues proches dans le cadre de UD. Comme nous l'avons montré dans l'article, une telle transformation peut poser divers problèmes, qui sont plus ou moins simples à résoudre. Il faut tout d'abord souligner le travail de réflexion qui est nécessaire en amont de la transformation. La transformation des corpus d'origine a été réalisée de manière essentiellement automatique et semi-automatique pour partie (nouveau découpage en tokens notamment). Ce travail sur des langues peu dotées n'aurait pu être réalisé sans une réelle coopération entre diverses équipes, dotées de compétences complémentaires, ce qui permet de gagner en efficacité. En effet, le travail en parallèle sur plusieurs langues a permis de profiter des expériences réalisées sur d'autres langues. Les problèmes qui se posent très vite sur une langue permettent une vigilance accrue à ce sujet dans une autre langue. Par ailleurs, les outils développés peuvent être réutilisés (scripts de conversion et de vérification notamment). Une première version de nos corpus au format CONLL-U est disponible (Bras *et al.*, 2018; Bernhard *et al.*, 2018a). Dans l'avenir, les informations manquantes seront également ajoutées : informations sur les propriétés lexicales et grammaticales des mots et relations syntaxiques. Pour l'occitan, l'annotation en dépendances est en cours dans le cadre du projet Linguatec.

Remerciements

Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet RESTAURE - référence ANR-14-CE24-0003).

5. http://universaldependencies.org/u/pos/PART.html

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In *Treebanks*, p. 165–187. Springer.

ALONSO H. M., SEDDAH D. & SAGOT B. (2016). From Noisy Questions to Minecraft Texts : Annotation Challenges in Extreme Syntax Scenario. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, p. 13–23.

ALONSO H. M. & ZEMAN D. (2016). Universal Dependencies for the AnCora treebanks. *Procesamiento del Lenguaje Natural*, (57), 91–98.

BENZITOUN C., FORT K. & SAGOT B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *JEP-TALN 2012 - Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles*, p. 99–112, Grenoble, France.

BERNHARD D., ERHART P., HUCK D. & STEIBLÉ L. (2018a). Annotated Corpus for the Alsatian Dialects. 10.5281/zenodo.2536041.

BERNHARD D., ERHART P., HUCK D. & STEIBLÉ L. (2018b). *Part-of-Speech Annotation Guidelines for the Alsatian Dialects*. 10.5281/zenodo.1171925.

BERNHARD D., LIGOZAT A.-L., MARTIN F., BRAS M., MAGISTRY P., VERGEZ-COURET M., STEIBLE L., ERHART P., HATHOUT N., HUCK D., REY C., REYNÉS P., ROSSET S., SIBILLE J. & LAVERGNE T. (2018c). Corpora with Part-of-Speech Annotations for Three Regional Languages of France : Alsatian, Occitan and Picard. In *11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan.

BRAS M. (2018). *Part-of-Speech Annotation Guidelines for the Occitan Language*. Rapport interne, UMR 5263 CLLE-ERSS, University of Toulouse. 10.5281/zenodo.1182949.

BRAS M., ESHER L., SIBILLE J. & VERGEZ-COURET M. (2018). *Annotated Corpus for Occitan*. Rapport interne, UMR 5263 CLLE-ERSS, University of Toulouse. 10.5281/zenodo.1182949.

BRAS M., VERGEZ-COURET M., HATHOUT N., SIBILLE J., SÉGUIER A. & DAZÉAS B. (2017). Loflòc : Lexic obert flechit occitan. In *XIIème Congrès de l'Association Internationale d'Etudes Occitanes*, Albi, France : Jean-François Courouau et al.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, p. 321–334.

CHUN J., HAN N.-R., HWANG J. D. & CHOI J. D. (2018). Building Universal Dependency Treebanks in Korean. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

GERDES K. & KAHANE S. (2016). Dependency annotation choices : Assessing theoretical and practical issues of universal dependencies. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, p. 131–140.

IDE N. & VÉRONIS J. (1994). Multext (Multilingual Tools and Corpora). In *14th Conference on Computational Linguistics (COLING'94), Kyoto, Japan*.

LIM K., PARTANEN N. & POIBEAU T. (2018). Analyse syntaxique de langues faiblement dotées à partir de plongements de mots multilingues. *Traitement Automatique des Langues*, **59**(3).

NIVRE J., DE MARNEFFE M.-C., GINTER F., GOLDBERG Y., HAJIC J., MANNING C. D., MCDONALD R., PETROV S., PYYSALO S., SILVEIRA N. & OTHERS (2016). Universal dependencies

v1 : A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

RAJMAN M., LECOMTE J. & PAROUBEK P. (1997). *Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique*. Rapport interne, EPFL & INaLF. GRACE GTR-3-2.1.

SANGUINETTI M., BOSCO C., LAVELLI A., MAZZEI A., ANTONELLI O. & TAMBURINI F. (2018). PoSTWITA-UD : an Italian Twitter Treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

SCHILLER A., TEUFEL S., STÖCKERT C. & THIELEN C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Rapport interne, Universität Stuttgart & Universität Tübingen.

SORIA C., MARIANI J. & ZOLI C. (2013). Dwarfs sitting on the giants' shoulders–how LTs for regional and minority languages can benefit from piggybacking major languages. In *Proceedings of XVII FEL Conference*, p. 73–79.

STRAKA M. & STRAKOVÁ J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 88–99, Vancouver, Canada.

TAULÉ M., MARTÍ M. A. & RECASENS M. (2008). Ancora : Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the 6th International Conference on Language Ressources and Evaluation (LREC2008)*, Marrakech, Morocco.

VERGEZ-COURET M. (2016). *Description du lexique Loflòc*. Research report, CLLE-ERSS.

VON REKOWSKI U. (1996). *ELM-FR : A typed French incarnation of the EAGLES-TS – Definition of Lexical Specification and Classification Guidelines*. Rapport interne, GSI-Erli.