Production automatique de gloses interlinéaires à travers un modèle probabiliste exploitant des alignements

Shu Okabe François Yvon
Université Paris-Saclay, CNRS, LISN, Bât. 508, Rue du Belvédère, F-91405 Orsay, France
{shu.okabe,francois.yvon}@limsi.fr

RÉSUMÉ

La production d'annotations linguistiques ou *gloses interlinéaires* explicitant le sens ou la fonction de chaque unité repérée dans un enregistrement source (ou dans sa transcription) est une étape importante du processus de documentation des langues. Ces gloses exigent une très grande expertise de la langue documentée et un travail d'annotation fastidieux. Notre étude s'intéresse à l'automatisation partielle de ce processus. Il s'appuie sur la partition des gloses en deux types : les gloses grammaticales exprimant une fonction grammaticale et les gloses lexicales indiquant les unités de sens. Notre approche repose sur l'hypothèse d'un alignement entre les gloses lexicales et une traduction ainsi que l'utilisation de Lost, un modèle probabiliste de traduction automatique. Nos expériences sur une langue en cours de documentation, le tsez, montrent que cet apprentissage est effectif même avec un faible nombre de phrases de supervision.

ABSTRACT

A Probabilistic Model for Automatic Interlinear Glossing and Alignment.

An important task in language documentation is the generation of linguistic annotations, or *interlinear glosses*), which express the meaning or role of each unit identified in a spoken source utterance (or its transcription). Glossing requires extensive expertise in the studied language and tedious work. In this article, we study ways to automate this process. Two types of glosses exist : grammatical glosses expressing a grammatical function and lexical glosses carrying the meaning of the unit. Our approach assumes that lexical glosses can be aligned with a target translation, enabling us to repurpose Lost, a probabilistic translation model for the glossing task. Our experiments on Tsez, a language in the process of being documented, show that useful glosses can be learned, even with a small number of supervision sentences.

MOTS-CLÉS : génération de gloses interlinéaires, documentation automatique des langues, alignement de mots.

KEYWORDS: interlinear gloss generation, computational language documentation, word alignment.

1 Introduction

Dans leur travail de documentation des langues, les linguistes de terrain produisent différentes strates d'annotations linguistiques des données orales collectées. Ces annotations permettent d'étudier plus en détail la langue et de préparer la production de dictionnaires ou de grammaires.

La figure 1 illustre ces différentes strates d'annotations appliquées à la transcription phonétique

S [source] :	sidaquł	yiła–r	ciq–q	allah–s	ašuni	b–ukad–n
G [gloses] :	one.day	DEM2.IISG.OBL–LAT	forest–POSS.ESS	God–GEN1	belt	III–see–PST.UNW
G' [catégories] :	LEX	GRAM–GRAM	LEX–GRAM	LEX–GRAM	LEX	GRAM–LEX–GRAM
T [traduction] :	*one day she saw a rainbow in the forest*					

FIGURE 1 – Exemple de strates d'annotation linguistique dans le cadre de la documentation de langue.

de l'enregistrement d'une phrase isolée. La phrase **S** dans la langue source étudiée (ici le tsez) est segmentée en deux niveaux : les unités lexicales sont séparées par des espaces et les morphèmes sont indiqués par les tirets au sein d'un mot. Une seconde étape d'annotation consiste à renseigner le sens ou la fonction grammaticale de chaque morphème. Ce niveau d'annotation est appelé *glose interlinéaire* (**G** sur la figure 1). Nous distinguons deux catégories de gloses sur la ligne **G'** : les gloses *grammaticales*, telles que LAT, indiquent la fonction du morphème ; les gloses *lexicales* (comme forest) expriment le sens du morphème, en utilisant un concept dans une langue de documentation (ici l'anglais). Enfin, une traduction **T** accompagne chaque phrase.

Si les phrases sources et leurs traductions peuvent être recueillies simultanément sur le terrain, elles ne sont glosées qu'ultérieurement lors d'étapes d'analyse. Elles sont ainsi coûteuses à obtenir et exigent une grande expertise et un fastidieux travail d'annotation manuelle. Il n'est alors pas surprenant d'observer que les ressources complètement annotées soient peu nombreuses au regard du volume d'enregistrements bruts (Seifart *et al.*, 2018). Notre objectif est ici d'étudier comment une partie de ce processus pourrait être automatisé, en effectuant une pré-annotation qui serait ensuite révisée par des annotateurs. En effet, les phénomènes complexes, qui, souvent, intéressent davantage les linguistes, ne sont pas les plus fréquents ; une automatisation pourrait être bénéfique pour traiter les cas les plus courants. Elle permettrait également d'améliorer la cohérence et la vitesse de l'annotation en gloses (Baldridge & Palmer, 2009). La tâche que nous considérons consiste alors à calculer des gloses (**G**) à partir de la phrase source segmentée en morphèmes (**S**) et de sa traduction (**T**).

Cette tâche soulève un grand nombre de difficultés, comme la faible quantité de phrases disponibles pour superviser cette annotation. De plus, si la variété de gloses grammaticales est en nombre fini (s'apparentant alors à une tâche d'étiquetage classique), les gloses lexicales sont, par nature, en nombre quasi illimité. Enfin, malgré certaines conventions partagées (comme les *Leipzig Glossing Rules* [1] (Bickel *et al.*, 2008)), ces annotations s'appuient sur une interprétation linguistique, toujours sujette à des variations inter- et intra-personnelles. Face à ces constats, la production automatique de gloses pour les langues en cours de documentation a été abordée de différentes manières, principalement sous l'angle d'un étiquetage de séquences impliquant plusieurs étapes (Samardžić *et al.*, 2015; Moeller & Hulden, 2018; Barriga Martínez *et al.*, 2021). Une approche classique repose sur l'utilisation de champs markoviens conditionnels (*Conditional Random Field*, CRF) (Lafferty *et al.*, 2001; Tellier & Tommasi, 2011). Par exemple, (McMillan-Major, 2020) emploie deux CRF : l'un pour prédire les gloses depuis la phrase source, un autre depuis la phrase traduite. Les deux prédictions sont ensuite combinées afin d'obtenir la prédiction finale. (Zhao *et al.*, 2020) met en œuvre une approche neuronale, considérant les deux entrées (phrases source et cible) comme séparées dans leur architecture. Ces approches illustrent les diverses solutions imaginées pour aborder le principal défi de la tâche de génération de gloses : le problème posé par l'inventaire des gloses lexicales.

Pour y répondre, nous supposerons qu'il est possible de dériver les gloses lexicales de la traduction.

1. Un inventaire en français des principales étiquettes grammaticales est proposé par B. Fradin (voir http://www.llf.cnrs.fr/fr/node/60).

Cette hypothèse permet de circonscrire l'ensemble des étiquetages d'une phrase donnée lors de l'entraînement et de l'inférence d'un modèle probabiliste, qui pourra alors prendre en charge un ensemble arbitraire d'étiquettes. Dans la section 2, après avoir formalisé la tâche, nous discutons du calcul de ces alignements, puis de leur considération dans un modèle de glose automatique. La section 3 présente les résultats obtenus sur le tsez, une langue déjà utilisée dans (Zhao *et al.*, 2020) [2].

2 Un modèle probabiliste pour les gloses interlinéaires

2.1 Formalisation du problème

Considérons tout d'abord la ligne **G'** de la figure 1. Son calcul peut se formaliser comme une tâche d'étiquetage de séquence classique, où chaque morphème source est associé à une étiquette binaire : LEX ou GRAM. Pour ce faire, une méthode standard est d'utiliser un CRF qui modélise :

$$p_\theta(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp\left\{\sum_{k=1}^{K}\theta_k G_k(\boldsymbol{x},\boldsymbol{y})\right\}}{\sum_{\boldsymbol{y'}\in\mathcal{Y}^T}\exp\left\{\sum_{k=1}^{K}\theta_k G_k(\boldsymbol{x},\boldsymbol{y'})\right\}} = \frac{1}{Z_\theta(\boldsymbol{x})}\exp\left\{\sum_{k=1}^{K}\theta_k G_k(\boldsymbol{x},\boldsymbol{y})\right\}, \quad (1)$$

où $\boldsymbol{x}$ représente la séquence de T morphèmes sources et $\boldsymbol{y}$ les étiquettes associées. Nous noterons $\mathcal{Y}$, l'ensemble des étiquettes possibles, ici restreint à deux possibilités $\mathcal{Y} = \{\text{LEX}, \text{GRAM}\}$. L'espace de recherche contenant l'ensemble des étiquetages possibles pour une phrase est alors égal à $\mathcal{Y}^T$. $\{G_k(\boldsymbol{x},\boldsymbol{y}), k \in [1,K]\}$ sont les caractéristiques et $\boldsymbol{\theta} \in \mathbb{R}^K$, le vecteur de paramètres associé. Une caractéristique teste des propriétés locales du couple $(\boldsymbol{x},\boldsymbol{y})$ à une position donnée (unigramme) ou à deux positions consécutives (bigramme). Enfin, $Z_\theta(\boldsymbol{x})$ est la fonction de partition qui normalise sur tous les étiquetages possibles et permet d'interpréter (1) comme une probabilité. L'apprentissage des paramètres à partir d'un corpus associant morphèmes et étiquettes est un processus standard et amplement documenté.

Il est possible d'étendre ce modèle pour prédire chaque glose *grammaticale* plutôt qu'une seule étiquette (GRAM). Ce changement implique un accroissement du nombre d'étiquettes possibles et de l'espace de recherche. Les calculs associés restent en effet réalisables même lorsque l'on considère plusieurs centaines d'étiquettes (Mueller *et al.*, 2013; Lavergne & Yvon, 2017). Cette méthode est utilisée dans des travaux précédents de prédiction de gloses par (Moeller & Hulden, 2018; Barriga Martínez *et al.*, 2021; Okabe & Yvon, 2022). Une fois les étiquettes grammaticales prédites par un CRF, les gloses lexicales (toutes regroupées sous l'étiquette LEX) sont annotées manuellement.

Calcul des gloses lexicales La prise en charge des gloses lexicales se heurte au problème de leur nombre et inventaire, non fixés au préalable. Deux hypothèses sont alors possibles :

H1 : considérer que seules les gloses lexicales observées à l'entraînement sont possibles, ce qui permet de spécifier *globalement* $\mathcal{Y}$, dont la taille pourra toutefois poser problème ;

2. Le code pour reproduire les expériences est disponible à l'adresse : https://github.com/shuokabe/gloss_lost.

H2 : supposer que les gloses lexicales peuvent également être déduites de la traduction, dans laquelle on peut s'attendre à retrouver les mêmes concepts évoqués. Cette hypothèse est notamment explorée par (McMillan-Major, 2020; Zhao *et al.*, 2020).

Sous l'hypothèse [H2], que nous adoptons également ici, gloser revient à prédire, pour chaque morphème, soit une étiquette grammaticale, soit un mot de la traduction (plus précisément, son lemme). [H1] et [H2] ne sont pas complètement exclusives et il est aussi possible d'inclure (certains) des mots observés à l'entraînement, même s'ils n'apparaissent pas dans la traduction. Ceci s'avère en particulier nécessaire, comme dans l'exemple 1, car les gloses lexicales comme God ou belt ne peuvent pas être déduites de la seule traduction (même en explorant les synonymes). L'utilisation d'un lexique complémentaire permet donc d'associer une glose lorsque celle-ci ne figure pas dans la traduction. La figure 2 illustre notre formalisation pour la phrase de l'exemple 1.

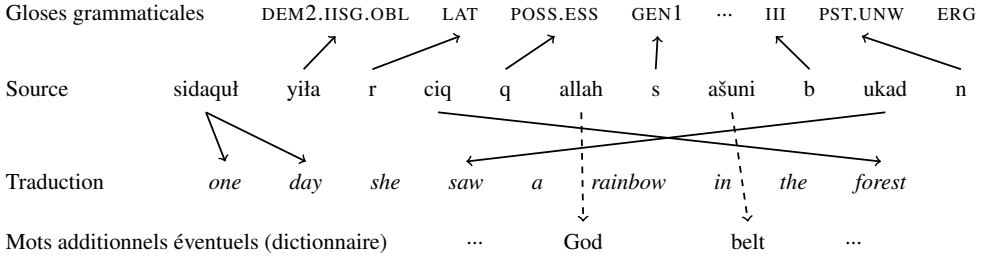


FIGURE 2 – Illustration du calcul de gloses pour la phrase de l'exemple 1. La partie supérieure contient les gloses grammaticales ; la partie inférieure, un alignement partiel entre la source et la traduction, ainsi que les mots supplémentaires (issus d'un dictionnaire).

Mettre en œuvre un modèle probabiliste inspiré de (1) demande de résoudre plusieurs problèmes : (a) définir les étiquetages possibles pour une phrase source donnée en exploitant les mots de la traduction et les gloses du corpus d'apprentissage, tout en prenant soin de contraindre l'espace de recherche associé pour que le calcul de Z_θ reste faisable ; (b) définir les étiquetages de référence pour superviser l'apprentissage, car les alignements entre source et traduction ne sont pas observés ; (c) celui de définir l'ensemble des étiquetages possibles à l'inférence.

Notre approche propose des réponses à chacun de ces problèmes. Pour le problème (a), nous utilisons une spécification *locale* (phrase par phrase) de $\mathcal{Y}$ incluant : les lemmes de la traduction, les gloses lexicales non-présentes dans la traduction, des associations fréquentes listées dans un dictionnaire. Pour ce qui concerne le problème (b), nous exploitons des alignements déterministes obtenus avec SimAlign (section 2.2) pour superviser l'apprentissage (cf. figure 2). Enfin, pour (c), nous augmentons à l'inférence l'espace de recherche avec d'autres mots-candidats, en nous basant sur un dictionnaire issu des données d'entraînement. Notre implémentation s'inspire de Lost (Lavergne *et al.*, 2011, 2013), un modèle probabiliste conçu pour la traduction automatique, qui permet de spécifier localement l'espace de recherche associé à un modèle globalement normalisé (section 2.3).

2.2 Supervision des alignements de gloses lexicales avec SimAlign

L'étape d'apprentissage du modèle CRF étendu demande de définir les étiquetages possibles pour chaque phrase, et au sein de cet ensemble, ceux qui sont jugés corrects. Selon [H2], nous supposons

que les étiquettes lexicales constituent un sous-ensemble des mots [3] apparaissant dans la traduction. Les étiquettes correctes (l'association entre morphèmes sources et mots de la traduction) ne sont pas observées : dans ce travail, nous les calculons de manière automatique en exploitant les gloses lexicales disponibles (voir exemple en figure 3) par alignement mot-à-mot entre glose et traduction.

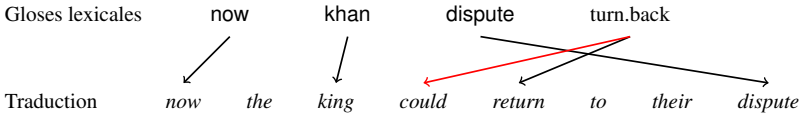
SimAlign L'alignement automatique de mots est réalisé par SimAlign (Jalili Sabet *et al.*, 2020), un modèle d'alignement neuronal basé sur la similarité entre plongements lexicaux de mots. SimAlign implante plusieurs méthodes pour obtenir un alignement à partir d'une matrice de similarité :

— `Argmax` aligne deux mots s'ils sont mutuellement plus proches voisins ;

— `Match` considère l'alignement comme un problème de couplage (*matching*) maximal dans le graphe biparti des mots sources et cibles, en utilisant les similarités pour pondérer les arêtes.

Ces deux méthodes produisent des alignements symétriques, dans lesquels chaque mot de la glose est associé au plus à un mot de la traduction (et réciproquement). Comme nous travaillons sur des gloses et des traductions en anglais, nous utilisons le modèle BERT anglais (BERT-base) (Devlin *et al.*, 2019) pour calculer les similarités. Nos expériences préliminaires montrent que les plongements lexicaux à la sortie de la couche 0 aboutissent au meilleur alignement glose-traduction, probablement dû à l'ordre très différent des gloses lexicales par rapport à l'anglais et l'absence de mots outils dans les gloses.

Comparaison des deux méthodes `Match` identifie par construction un alignement pour chaque glose lexicale [4] et conduit donc à un score de rappel élevé. En comparaison, `Argmax` propose moins de liens d'alignement, mais avec une meilleure précision. Des statistiques sur les alignements ainsi obtenus sont données en section 3.1.



source	howži	xan	yizi–z	daʕba–ƛ'or	uti–ɬ–n
référence	now	**khan**	DEM2.IPL.OBL–GEN2	dispute–SUPER.LAT	**turn.back**–POT–PST.UNW
`Match`	now	**king**	DEM2.IPL.OBL–GEN2	dispute–SUPER.LAT	**could.return**–POT–PST.UNW
`Argmax`	now	**king**	DEM2.IPL.OBL–GEN2	dispute–SUPER.LAT	**return**–POT–PST.UNW

FIGURE 3 – Exemple d'alignements calculés par SimAlign (haut ; heuristique `Argmax` en noir, rouge et noir pour `Match`) et d'étiquettes obtenues en les utilisant (bas). Les différences sont en **gras**.

La figure 3 présente un exemple de phrase étiquetée en utilisant les alignements obtenus par ces deux méthodes. La différence principale réside dans les deux gloses lexicales khan et turn.back qui ne sont pas présentes dans la traduction. Pour la première, SimAlign trouve correctement l'association entre khan et king, ce qui permet d'obtenir une étiquette de sens similaire depuis la traduction. Dans le second cas, `Match` identifie deux liens pour turn.back [5], alors que `Argmax` n'en identifie qu'un.

3. L'alignement est effectué avec la traduction telle quelle, mais les étiquettes associées correspondent aux lemmes des mots identifiés par alignement.

4. Sauf quand la traduction comporte moins d'unités que de gloses lexicales (soit 31 phrases dans nos expériences).

5. Lors de l'alignement, le « . » est enlevé, d'où les deux liens d'alignement. Ce choix nous a permis d'obtenir de meilleurs résultats d'alignements dans des expériences préliminaires.

Cas des gloses non-alignées Dans les cas où SimAlign n'identifie pas de liens d'alignement pour une glose, nous attribuons une étiquette dédiée (« unk »). Cette dernière pouvant concerner une proportion non-négligeable dans le corpus, en particulier dans le cas de `Argmax` (voir section 3.1), nous complétons ces alignements en exploitant un dictionnaire. Ce dernier associe à tout morphème *source*, l'étiquette lexicale qui lui est la plus souvent associée dans la base d'entraînement. Ce « lexique » est ensuite utilisé pour éventuellement attribuer une étiquette lorsque l'alignement fait défaut. Nous assignons ainsi aux morphèmes soit les lemmes des mots de la traduction, soit, en l'absence d'alignement, des étiquettes supplémentaires obtenues via le dictionnaire (cf. figure 2).

2.3 Mise en œuvre de Lost pour la prédiction de gloses

Notre implémentation du modèle probabiliste pour la prédiction de gloses repose sur le système Lost (Lavergne *et al.*, 2011, 2013) conçu originellement dans un cadre de traduction statistique.

Lost est un modèle de traduction à base de segments qui étend les CRF de manière à prendre en charge de très grands ensembles d'étiquettes et de caractéristiques, afin de pouvoir associer aux segments sources des étiquettes correspondant à leurs traductions en langue cible. Dans cette implémentation, il est possible de circonscrire l'ensemble des étiquetages possibles pour chaque phrase, sur la base d'un ensemble réduit d'étiquettes pour chaque mot à annoter. Limiter l'espace de recherche rend l'apprentissage computationnellement faisable en simplifiant le calcul de $Z_\theta(\boldsymbol{x})$ dans l'équation (1).

Espace de recherche Selon [H2], l'espace de recherche est restreint à l'ensemble des étiquettes grammaticales observées à l'entraînement, complété par des lemmes des mots dans la traduction de la phrase traitée (cf. figure 2). À l'apprentissage, nous ajoutons à cet ensemble les gloses des morphèmes non-alignés pour que toute phrase de référence soit atteignable ; à l'inférence, les gloses de référence sont inconnues, et nous ajoutons pour chaque morphème connu l'étiquette lexicale la plus fréquemment associée dans l'ensemble d'apprentissage, conduisant à mettre en œuvre un croisement des hypothèses [H1] & [H2].

Étiquettes simples Notre première configuration utilise la partie supérieure des caractéristiques présentées dans le tableau 1. Cette configuration est illustrée dans la figure 4 : en entrée, Lost reçoit pour chaque morphème source m, sa position t au sein du mot et sa longueur l ; en sortie le modèle prédit uniquement la glose g, supervisée par les liens d'alignement (section 2.2).

Étiquettes structurées La configuration avec étiquettes structurées enrichit la représentation de la sortie de deux informations : d'une part, une étiquette binaire b (GRAM ou LEX) indiquant la nature de la glose, d'autre part, l'étiquette en partie du discours (PoS) p associée au mot aligné dans la phrase traduite. Ces deux informations, qui dérivent de manière déterministe de l'étiquette de base, permettent de construire des caractéristiques supplémentaires (cf. partie inférieure du tableau 1) dont l'estimation est plus robuste [6]. Elles permettent également de généraliser l'étiquetage à des morphèmes inconnus.

6. Pour les étiquettes lexicales issues du dictionnaire, nous utilisons le PoS qui lui est le plus fréquemment associé dans le corpus d'apprentissage.

Caractéristique	Test	Exemple (cf. figure 4, $i = 4$))
uni-gloss	$\mathbb{1}(g_i = g)$	PST.UNW
bi-gloss	$\mathbb{1}(g_i = g) \wedge \mathbb{1}(g_{i-1} = g')$	(say, PST.UNW)
uni-gloss-morph	$\mathbb{1}(g_i = g) \wedge \mathbb{1}(m_i = m)$	n
uni-gloss-position	$\mathbb{1}(g_i = g) \wedge \mathbb{1}(t_i = t)$	1
uni-gloss-length	$\mathbb{1}(g_i = g) \wedge \mathbb{1}(l_i = l)$	1
bi-gloss-morph	$\mathbb{1}(g_i = g) \wedge \mathbb{1}(g_{i-1} = g') \wedge \mathbb{1}(m_i = m)$	(say, PST.UNW, n)
uni/bi-bin	$\mathbb{1}(b_i = b)$ $(\wedge \mathbb{1}(b_{i-1} = b'))$	GRAM ((LEX, GRAM))
uni/bi-pos	$\mathbb{1}(p_i = p)$ $(\wedge \mathbb{1}(p_{i-1} = p'))$	GRAM ((VERB, GRAM))
uni-bin-morph/position/length	$\mathbb{1}(b_i = b) \wedge \mathbb{1}(m_i = m) / \mathbb{1}(t_i = t) / \mathbb{1}(l_i = l)$	(GRAM, n) / (GRAM, 1) / (GRAM, 1)
bi-bin-gloss	$\mathbb{1}(g_i = g) \wedge \mathbb{1}(b_{i-1} = b')$	(LEX, PST.UNW)
bi-gloss-bin	$\mathbb{1}(b_i = b) \wedge \mathbb{1}(g_{i-1} = g')$	(say, GRAM)
uni-pos-morph	$\mathbb{1}(p_i = p) \wedge \mathbb{1}(g_i = g)$	(GRAM, n)
bi-pos-gloss	$\mathbb{1}(g_i = g) \wedge \mathbb{1}(p_{i-1} = p')$	(VERB, PST.UNW)
bi-gloss-pos	$\mathbb{1}(p_i = p) \wedge \mathbb{1}(g_{i-1} = g')$	(say, GRAM)

TABLE 1 – Caractéristiques unigrammes et bigrammes pour le modèle avec étiquettes simples (haut) et structurées (haut + bas).

i	Entrées			Sorties			Référence
	morphème source m	position (mot) t	longueur l	glose approximée g	GRAM ou LEX b	étiquette PoS p	glose de référence
0	q'orol	0	6	widow	LEX	NOUN	widowed
1	ɣˤanabi	0	7	the	LEX	DET	woman
2	a	1	1	ERG	GRAM	GRAM	ERG
3	eʔi	0	3	say	LEX	VERB	say
4	n	1	1	PST.UNW	GRAM	GRAM	PST.UNW

FIGURE 4 – Exemple de configuration pour « q'orol ɣˤanabi–a eʔi–n » (*the widow said*). L'espace de recherche est défini localement par $\mathcal{Y} = \{\text{étiquettes grammaticales}\} \bigcup \{\text{the, widow, say, unk}\}$.

2.4 Conditions expérimentales

Langue Nous étudions principalement le tsez, une langue nakho-daghestanienne parlée dans la république du Daghestan en Russie. Elle est actuellement en cours de documentation (Comrie & Polinsky, à paraître).

Le corpus sur lequel nous travaillons est constitué de récits du folklore, entièrement glosés et traduits en anglais et en russe (Abdulaev & Abdulaev, 2010). Il est constitué de 2 000 phrases avec 40 229 occurrences de morphèmes et 1 603 types de morphèmes (ce qui correspond à 20 153 occurrences de mots). Il existe 157 gloses grammaticales représentant 54 % des étiquettes du corpus et 1 363 gloses lexicales. Dans la mesure où les étiquettes lexicales correspondent à des lemmes en langue anglaise, nous utilisons les traductions en anglais pour calculer les gloses. Notons toutefois que le modèle d'alignement est lui multilingue et pourrait s'appuyer sur les traductions en russe. Ce corpus a, par ailleurs, déjà été étudié dans (Zhao *et al.*, 2020) pour la même tâche [7]. Le tsez fait également partie des langues considérées par le défi organisé en 2023 et visant à évaluer le calcul automatique de gloses interlinéaires [8].

7. Les auteurs ayant un autre découpage du texte (1 782 phrases au total), leurs résultats ne se comparent pas directement aux nôtres. À titre indicatif, ces auteurs obtiennent une correction de 87 avec leurs modèles neuronaux et le modèle à base de CRF de (McMillan-Major, 2020) utilisé comme *baseline*, obtient un score de 84.

8. https://github.com/sigmorphon/2023GlossingST.

Nous divisons le corpus en trois sous-parties cohérentes : 200 phrases pour les jeux de développement et de test, et un jeu d'entraînement de taille variable (200, 500, 1 000 et 1 600), ce qui nous permet d'étudier l'évolution des performances du modèle en fonction des données disponibles. Nous conservons l'ordre des phrases dans le corpus afin de préserver la cohérence des récits.

Métrique Nous donnerons principalement le taux de correction des étiquettes, à savoir la proportion d'étiquettes correctement prédites. Dans nos analyses, nous nous intéressons aussi au rappel différencié selon la nature du morphème (grammaticale ou lexicale).

Paramétrage Nous avons utilisé les paramètres par défaut de Lost. Si Lost permet une régularisation *elastic net*, nous n'utilisons que la régularisation l_1 (paramétrage par défaut avec un poids de $0{,}5$), qui, d'après les expériences préliminaires, semble suffisante. Nous stoppons l'apprentissage après 15 itérations complètes. Pour la lemmatisation et la génération des étiquettes de PoS des traductions en anglais, nous utilisons spaCy [9].

Baseline Le modèle de référence (`maj`) affecte l'étiquette majoritaire observée dans la base d'apprentissage. Comme il n'utilise que les associations entre morphème source et glose déjà vus, il reproduit en quelque sorte le système de « lexique » présent dans certains outils d'annotation tels que ELAN-CorpA ou FieldWorks Language Explorer (FLEx) (Rogers, 2010).

3 Résultats expérimentaux

Dans cette section, nous cherchons tout d'abord à vérifier la validité de l'hypothèse [H2], à savoir la présence des gloses lexicales dans la traduction (section 3.1). Ensuite, nous comparons les performances des configurations, de manière générale et séparée sur les deux catégories de gloses, puis évaluons l'impact de l'alignement et du dictionnaire sur les prédictions (section 3.2).

3.1 Statistiques sur les étiquettes obtenues par alignement

Nous étudions tout d'abord les étiquettes obtenues par alignement. Quelle proportion de gloses lexicales parvient-on à aligner automatiquement ? De plus, nous supposons que les gloses lexicales peuvent être retrouvées dans la traduction : dans quelle mesure est-ce vérifié dans nos projections d'alignements ?

Couverture des gloses obtenues par alignement Le tableau 2 présente le nombre (d'occurrences) et la proportion [10] d'étiquettes lexicales non-alignées avec les deux méthodes de SimAlign. La méthode `Match`, par construction, aligne (presque) toutes les gloses lexicales ; l'apport d'un lexique y est donc limité. En revanche, pour la méthode `Argmax`, qui laisse non-alignées près de 20 % de gloses lexicales, l'utilisation du dictionnaire est essentielle et permet d'aligner plus de 95% des morphèmes dès que l'on dispose de 500 phrases d'entraînement.

9. https://spacy.io/, *pipeline* `en_core_web_sm`.
10. Il y a 18,635 gloses lexicales au total.

Taille du corpus	base /	+ dictionnaire 200	500	1 000	1 600
`Argmax`	3 615 (19,4 %)	1 223 (6,6 %)	858 (4,6 %)	733 (3,9 %)	627 (3,4 %)
`Match`	35 (0,2 %)	18 (0,1 %)	16 (0,1 %)	9 (0,0 %)	9 (0,0 %)

TABLE 2 – Occurrences (et proportion) de gloses lexicales non-alignées par SimAlign et éventuellement complétées par un dictionnaire créé à partir d'un corpus d'apprentissage.

Comparaison avec les gloses de référence Le tableau 3 présente la proportion d'étiquettes obtenues par alignement qui correspondent exactement aux gloses de référence : 80 % des gloses de référence sont prédictibles directement en se basant sur des projections d'alignement depuis la traduction. Comme nous mesurons des correspondances exactes, il est probable que ce taux serait plus élevé si l'on tenait compte des proximités sémantiques (comme khan et king dans l'exemple de la figure 3). Ce premier résultat conforte l'hypothèse [H2].

Dans le cas de base (sans dictionnaire), la méthode `Match` obtient bien une valeur bien plus élevée qu'`Argmax`. La différence entre les méthodes reste toutefois faible au regard du nombre d'alignements supplémentaires identifiés par `Match` (cf. tableau 2), qui correspondent souvent à des erreurs (the / woman figure 4) plutôt qu'à des mots sémantiquement proches.

En utilisant le dictionnaire créé avec les données d'entraînement, toutefois, la tendance s'inverse : si la qualité des étiquettes stagne avec `Match`, elle s'améliore significativement avec `Argmax`. Bien que certains morphèmes restent non-alignés, l'emploi d'un dictionnaire permet néanmoins de se rapprocher des données de référence, tout en restant compatible avec les conditions du test. Dans notre approche, il reste toujours possible que certains morphèmes reçoivent l'étiquette conventionnelle « unk » à l'apprentissage ou au test, leur désambiguïsation restant alors à la charge d'un annotateur.

Taille du corpus	base /	+ dictionnaire 200	500	1 000	1 600
`Argmax`	80,7	84,3	85,1	85,0	85,2
`Match`	81,5	81,5	81,5	81,5	81,6

TABLE 3 – Correspondances exactes entre les gloses de référence et celles obtenues par alignement.

3.2 Résultats de l'étiquetage

Le tableau 4 présente le score de correction des différentes configurations présentées en section 2.3 et le tableau 5 détaille le rappel pour les deux catégories de gloses. Nous remarquons tout d'abord que la *baseline* `maj` est dans nos conditions expérimentales, un modèle d'étiquetage compétitif, avec les meilleurs résultats au niveau des gloses lexicales, pour toute taille de corpus d'entraînement.

L'utilisation du modèle probabiliste permet toutefois de mieux prédire les gloses grammaticales, qui sont plus ambiguës, et s'avère globalement meilleur que la *baseline* à partir de 1 000 phrases environ. Concernant Lost, les configurations avec dictionnaire s'avèrent toujours meilleures que leur équivalent sans dictionnaire. Ceci illustre le bénéfice de l'augmentation de l'espace de recherche par des lemmes probables (d'après le morphème source) mais absents de la traduction, comme en témoigne également la hausse du rappel pour les gloses lexicales (cf. tableau 5). L'utilisation d'étiquettes structurées

Configuration		maj	base simple	base struct.	dict simple	dict struct.
Argmax	200	**64,1**	54,0	53,9	57,7	59,2
	500	**72,5**	62,4	63,5	69,6	71,3
	1000	74,9	67,3	68,3	73,6	**75,3**
	1600	77,1	69,6	71,1	77,1	**77,8**
Match	200	**64,1**	55,5	56,3	59,2	59,7
	500	**72,5**	64,1	65,8	69,3	71,0
	1000	74,9	68,6	69,5	74,1	**76,0**
	1600	77,1	71,0	72,0	77,4	**78,1**

TABLE 4 – Évolution de la correction selon la taille des données d'entraînement. Pour comparer avec chaque méthode d'alignement, les résultats de maj (qui lui n'en dépend pas) sont répétés.

Configuration		P^0_{lex}	maj gram	maj lex	base simple gram	base simple lex	base structuré gram	base structuré lex	dict simple gram	dict simple lex	dict structuré gram	dict structuré lex
Argmax	200	32,3 %	71,2	**56,1**	**81,6**	23,4	80,0	25,1	78,5	34,7	79,8	36,4
	500	16,9 %	72,4	**72,7**	85,8	36,5	**86,5**	38,0	83,9	53,8	86,3	54,8
	1000	9,9 %	72,8	**77,3**	89,2	42,9	89,4	45,0	88,7	56,9	**90,0**	59,1
	1600	4,5 %	73,0	**81,6**	90,5	46,5	**91,0**	49,0	90,1	62,7	90,8	63,5
Match	200	32,3 %	71,2	**56,1**	79,3	29,1	79,2	30,9	**80,5**	35,6	80,1	37,2
	500	16,9 %	72,4	**72,7**	84,9	41,1	86,4	43,1	85,0	51,9	**87,0**	53,3
	1000	9,9 %	72,8	**77,3**	88,2	46,9	88,8	48,0	88,2	58,5	**89,2**	61,3
	1600	4,5 %	73,0	**81,6**	89,2	50,8	**90,3**	51,6	88,9	64,7	89,8	65,1

TABLE 5 – Scores de rappel différenciés selon la nature de la glose. P^0_{lex} indique la proportion d'étiquettes *lexicales* présentes à l'*inférence* mais jamais observées à l'*entraînement*.

améliore sensiblement les prédictions, illustrant l'intérêt des caractéristiques supplémentaires faisant intervenir des catégories plus générales (et robustes) comme GRAM, LEX ou les PoS. Enfin les deux méthodes d'alignement de SimAlign sont très proches, avec un petit avantage pour Match. Une explication est que cette méthode génère moins d'étiquettes unk qu'Argmax ; par exemple, pour la configuration utilisant des étiquettes structurées et un dictionnaire, le premier produit environ 3 % d'étiquettes inconnues, contre 7 % pour le second.

Il faut finalement rappeler que les résultats du tableau 4 évaluent les correspondances *exactes* entre les gloses de référence et les prédictions et sous-estiment la qualité des gloses automatiques. Ceci est illustré par la figure 3 où les modèles basés sur Lost proposent king, pourtant comptabilisé comme une erreur car différant de la référence (kahn).

4 Conclusion

Dans cet article, nous avons présenté une nouvelle approche pour aborder la tâche de génération automatique de gloses linéaires. D'une part, après avoir validé l'hypothèse que les gloses lexicales sont le plus souvent présentes dans les traductions, nous avons utilisé des alignements automatiques pour superviser l'apprentissage d'un modèle de glose, en complétant éventuellement avec un dictionnaire. D'autre part, nous avons eu recours à une extension du modèle CRF, Lost, permettant de restreindre

et sélectionner les gloses lexicales possibles pour une phrase donnée. Dans nos conditions expérimentales, nous parvenons alors à surpasser le modèle de base avec 1 000 phrases d'entraînement, l'apprentissage améliorant en particulier les gloses grammaticales.

Ces résultats sous-estiment la qualité des gloses automatiques, du fait de la mesure considérée : pour quantifier cette sous-estimation, nous prévoyons de réaliser un alignement manuel entre source et traduction, afin de disposer de références qui correspondent exactement au problème d'apprentissage traité. Pour le futur, plusieurs pistes sont explorées, en particulier l'utilisation des jeux de caractéristiques plus riches et l'exploration des manières alternatives de construire l'espace de recherche pour l'étiquetage. Il serait par exemple intéressant d'exploiter également des alignements entre mots/morphèmes grammaticaux (en source et en cible) qui pourraient aider à mieux localiser les morphèmes pleins qui leur sont proches. Pour améliorer l'apprentissage, il est aussi envisagé de relâcher les contraintes découlant de l'utilisation d'un alignement (ici, calculé par SimAlign) et d'apprendre le modèle probabiliste avec des alignements latents. Une autre perspective est d'augmenter la base de données utilisées à l'apprentissage en combinant des corpus documentant plusieurs langues : l'objectif sera alors de faire émerger des caractéristiques « multilingues » dans les langues documentées pour améliorer la robustesse de caractéristiques testant des propriétés génériques (par exemple, la longueur ou la position relative des morphèmes grammaticaux vs. non-grammaticaux).

Remerciements

Ce travail est effectué dans le cadre du projet franco-allemand ANR-DFG « La documentation automatique des langues à l'horizon 2025 » (*Computational Language Documentation by 2025*, CLD 2025, ANR-19-CE38-0015-04). Les auteurs remercient chaleureusement Thomas Lavergne pour l'accompagnement dans les expériences avec Lost, ainsi qu'Antonios Anastasopoulos pour la mise à disposition du corpus tsez.

Références

ABDULAEV A. K. & ABDULAEV I. K. (2010). *Cezjas fol'klor : (gíurus mecrek°iorno butirno) = Dido (Tsez) folklore = Didojskij (cezskij) fol'klor*. Leipzig : Lotos.

BALDRIDGE J. & PALMER A. (2009). How well does active learning *actually* work ? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, p. 296–305, Singapore : Association for Computational Linguistics.

BARRIGA MARTÍNEZ D., MIJANGOS V. & GUTIERREZ-VASQUES X. (2021). Automatic interlinear glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, p. 34–43, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.americasnlp-1.5.

BICKEL B., COMRIE B. & HASPELMATH M. (2008). The Leipzig Glossing Rules : Conventions for interlinear morpheme-by-morpheme glosses. Leipzig : Max Planck Institute for Evolutionary Anthropology, Department of Linguistics. https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

COMRIE B. & POLINSKY M. (à paraître). Tsez. In Yuri Koryakov, Yury Lander and Timur Maisak (eds.) The Caucasian Languages. An International Handbook. Mouton. HSK series.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

JALILI SABET M., DUFTER P., YVON F. & SCHÜTZE H. (2020). SimAlign : High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 1627–1643, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.147.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LAVERGNE T., ALLAUZEN A., CREGO J. M. & YVON F. (2011). From n-gram-based to CRF-based translation models. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, p. 542–553, Edinburgh, Scotland : Association for Computational Linguistics.

LAVERGNE T., ALLAUZEN A. & YVON F. (2013). Un cadre d'apprentissage intégralement discriminant pour la traduction statistique. In *Actes de la 20ème Conférence sur le Traitement Automatique des Langues Naturelles*, p. 450–463, Les Sables d'Olonne, France : ATALA. HAL : hal-01908381.

LAVERGNE T. & YVON F. (2017). Learning the structure of variable-order CRFs : a finite-state perspective. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2017, p. 433–439 : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1044.

MCMILLAN-MAJOR A. (2020). Automating gloss generation in interlinear glossed text. In *Proceedings of the Society for Computation in Linguistics 2020*, p. 355–366, New York, New York : Association for Computational Linguistics.

MOELLER S. & HULDEN M. (2018). Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, p. 84–93, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

MUELLER T., SCHMID H. & SCHÜTZE H. (2013). Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 322–332, Seattle, Washington, USA : Association for Computational Linguistics.

OKABE S. & YVON F. (2022). Vers la génération automatique de gloses pour la documentation automatique des langues. In L. BECERRA, B. FAVRE, C. GARDENT & Y. PARMENTIER, Éds., *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, p. 198–203, Marseille, France : CNRS. HAL : hal-03846843.

ROGERS C. (2010). Review of Fieldworks Language Explorer (FLEx) 3.0. In *Language Documentation & Conservation 4*, p. 78–84.

SAMARDŽIĆ T., SCHIKOWSKI R. & STOLL S. (2015). Automatic interlinear glossing as two-level sequence classification. In *Proceedings of the 9th SIGHUM Workshop on Language Technology*

for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), p. 68–72, Beijing, China : Association for Computational Linguistics. DOI : 10.18653/v1/W15-3710.

SEIFART F., EVANS N., HAMMARSTRÖM H. & LEVINSON S. (2018). Language documentation twenty-five years on. *Language*, **94**(4), e324–e345. DOI : 10.1353/lan.2018.0070.

TELLIER I. & TOMMASI M. (2011). Champs Markoviens Conditionnels pour l'extraction d'information. In E. GAUSSIER & F. YVON, Éds., *Modèles probabilistes pour l'accès à l'information textuelle*, p. 223–267. Hermès. HAL : inria-00514525.

ZHAO X., OZAKI S., ANASTASOPOULOS A., NEUBIG G. & LEVIN L. (2020). Automatic interlinear glossing for under-resourced languages leveraging translations. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 5397–5408, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.471.