Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Corpus multilingue annoté pour l'étude sémantique des expressions quantifiantes – Problèmes de segmentation du coréen et du japonais

Raoul Blin[1] Jinnam Choi[2]
(1) CNRS-CRLAO, Campus Condorcet Bât. Recherche Sud 5 Cours des Humanités, 93322 Aubervilliers, France
(2) CLLE, Université Jean-Jaurès, allées A.Machado, 31058 Toulouse cedex 9, France
blin@ehess.fr, jinnamchoi11@gmail.com

RÉSUMÉ

Le travail présenté dans cet article s'inscrit dans le projet de constitution d'un corpus comparable, annoté pour l'étude sémantique de la quantification en coréen, français, japonais et chinois mandarin. Le corpus est annoté en dépendances au format SUD. Nous montrons la nécessité d'adopter une segmentation plus fine que celle en usage habituellement pour le coréen et le japonais. Cette segmentation améliore la description de la quantification dans environ 5% des phrases par rapport à la segmentation usuelle. Elle permet aussi une analyse morpho-syntaxique plus fine.

ABSTRACT

Annotated multilingual corpus for the semantic study of quantifying expressions – Segmentation problems in Korean and Japanese

The study presented in this article is part of a project to develop a comparable, annotated corpus for the semantic analysis of quantification in Korean, French, Japanese and Mandarin Chinese. The corpus is annotated with dependencies in SUD format. We demonstrate the necessity of adopting a finer segmentation system for Korean and Japanese than that conventionally used. This enhanced segmentation improves the treatment of quantifying expressions in approximately 5% of sentences compared to the usual segmentation, and enables a finer morpho-syntactic analysis.

MOTS-CLÉS : corpus comparable, quantification, coréen, japonais, SUD.

KEYWORDS : comparable corpus, quantification, Korean, Japanese, SUD.

1 Introduction

La quantification du nom est un des grands domaines de recherche en sémantique linguistique. Par quantification, nous entendons ici l'expression de la définitude, de la quantité et de la distributivité. Les trois sont souvent exprimées par les mêmes expressions et sont ainsi intimement liées. Par exemple en français, les articles *le/la/les* expriment le défini ou le générique, et la quantité (singulier ou pluriel). Les numéraux (ex : *un, deux,...*) expriment la quantité et l'indéfini. Le distributif *chaque* exprime aussi le nombre puisqu'il

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 196–205.

«distribue» le prédicat sur les individus d'un ensemble.

La quantification est présente dans les langues plus ou moins explicitement, et sous différentes formes. Ainsi l'anglais ou encore les langues romanes, explicitent systématiquement la quantification des noms communs, essentiellement avec un déterminant. A l'inverse, des langues comme le chinois (mandarin), le coréen, ou le japonais, l'explicitent rarement. Elles disposent toutefois d'outils pour le faire, par exemple des déterminants ou des affixes. Cette différence entre langues est bien connue par exemple des traducteurs, qui doivent choisir un déterminant lorsque la langue cible est à déterminant obligatoire et que la langue source ne fournit pas d'indication sur la quantification (voir (Bond, 2001) pour le japonais par exemple). Cette différence est aussi mentionnée en acquisition du langage (par exemple (Marsden, 2009) parmi beaucoup d'autres.).

La comparaison inter-langues des modes de quantification des expressions nominales (noms ou morphèmes à caractère nominal) reste à notre connaissance peu documentée. Par exemple Chierchia (1998) propose une typologie des noms nus et donc indirectement de la quantification, mais ne fournit pas de données à l'appui. D'autres ne traitent que quelques quantificateurs (ex : Kuno *et al.* 1999) et se concentrent en général sur une comparaison à l'anglais. D'un autre côté, les descriptions multilingues à grande échelle ne fournissent pas d'informations spécifiques sur la quantification. C'est le cas par exemple des treebanks annotés en dépendances UD ou SUD [1] .

Pour subvenir aux besoins en données, nous constituons actuellement un corpus de grande taille, comparable et annoté pour quatre langues : chinois (mandarin), coréen (pratiqué en Corée du Sud), français et japonais. Pour l'étude sémantique de la quantification, l'annotation doit fournir (au moins) trois informations : description des expressions quantifiantes, des expressions nominales quantifiables, des éventuelles dépendances entre ces expressions.

La stratégie que nous adoptons consiste à annoter le corpus en dépendances au format SUD (Gerdes *et al.*, 2019) et à y ajouter une couche sémantique, en s'appuyant entre autres sur les informations syntaxiques. Le choix s'est porté sur une annotation en dépendances car elle est réputée être efficace en traitement automatique. Le format SUD en particulier a été retenu car il est explicitement orienté syntaxe, qu'il a été éprouvé sur de grands corpus et dans de très nombreuses langues dont celles traitées ici.

Néanmoins, un travail exploratoire a montré que pour le coréen et le japonais, une partie des expressions étudiées étaient incorrectement décrites par l'annotation au format SUD. Il s'agit des expressions qui apparaissent dans les mots d'origine chinoise. La faute revient à la segmentation. Par exemple, le «token» (co) gakkuk / (ja) 各国 *kakkoku* («chaque pays») est en fait constitué d'un quantificateur (co) *gak* / (ja) *kaku* et d'un morphème nominal (co) *guk* / (ja) *koku* «pays». Mais le tout étant décrit comme un token unique, ses constituants et leur relation de dépendance échappe aux descriptions. De même, un token comme (co) *geonsu* / (ja) 件数 *kensū* «nombre de cas» est constitué d'un nom («nombre») et de son argument nominal («cas»). Faute de segmentation, il n'est pas possible de signaler le fait que l'argument n'est pas quantifié et que cela rejoint ce qui s'observe en français au niveau syntaxique : *nombre de ∅ cas*.

1. Dans cet article, pour tout ce qui concerne l'annotation au format (S)UD, nous ferons exclusivement référence aux treebanks disponibles respectivement sur universaldependencies.org et surfacesyntacticud.org

Nous nous concentrons dans le présent article sur le coréen et le japonais. Nous présentons la méthode d'annotation des informations sémantiques relatives à la quantification, et discutons d'une segmentation adaptée à ces deux langues, y compris pour le vocabulaire sino-coréen et sino-japonais (désormais «sino-CJ»). En section 2 nous décrivons de façon non exhaustive les modes de quantification en coréen et japonais et dans la section suivante 3 l'annotation que nous adoptons pour la quantification. Dans la section 4, nous expliquons pourquoi la segmentation habituellement utilisée en SUD dégrade la qualité des descriptions de ces deux langues, en particulier pour ce qui concerne le vocabulaire d'origine chinoise. Nous proposons donc une nouvelle segmentation (section 5) et adaptons l'annotation syntaxique et relative à la quantification (sections 6). Nous proposons enfin en section 7 une expérimentation sur le japonais pour montrer comment sont implémentées la segmentation et l'annotation. Nous mesurons quelle quantité d'information cette combinaison permet de récupérer.

2 Les expressions quantifiantes en coréen et japonais

Nous proposons une brève synthèse des modes de quantification des expressions nominales quantifiables en coréen et japonais. Les expressions sont des syntagmes nominaux ou des compositions de morphèmes nominaux. Pour simplifier, nous parlerons aussi de «nom (commun)».

Le coréen et le japonais sont des langues SOV. Elles marquent les cas à l'aide de particules casuelles. Le nom commun peut à lui seul constituer un argument nominal, même en l'absence de déterminant et de toute autre marque de quantification. On parle alors de «nom nu». Dans le discours, la majorité des occurrences de noms sont nues (un travail en cours, à paraître, montre que dans du texte encyclopédique et journalistique, 70% des noms sont nus). Les deux langues ne sont pas pour autant dépourvues d'expressions quantifiantes. Dans l'exemple suivant, (ja) *neko* et *nezumi* sont nus (sans marque de quantification), *gakusei* est quantifié par un démonstratif : il est défini, de nombre indéterminé. Le nom commun est invariable pour le nombre et tout autre type de quantification. L'analyse est identique pour la traduction en coréen : geu_{DEM} $haksaeng_{\text{NC}}$ ui_{GEN} *goyangi* $neun_{\text{TH}}$ *jwi* $reul_{\text{OBJ}}$ $meogeossda_{\text{V}}$.

(1) (ja) *sono* gakusei no *neko wa* *nezumi wo* *tabete-shimatta.*
DEM NC GEN chat TH/SUJ souris OBJ avoir-mangé.
le/s chat/s de cet étudiant a/ont mangé des/les/la souris.

Contrairement au français pour lequel la quantification se fait essentiellement par le déterminant [2], les expressions quantifiantes sont variées en japonais et coréen. Le tableau 1 propose une liste non exhaustive, pour le japonais. Y sont indiquées à titre indicatif la partie de discours (SUD-POS) et la nature de la dépendance (SUD-dep) au nom quantifié. Nous fournissons aussi les parties du discours selon l'UniDic (Den *et al.*, 2007) indiquées dans (XPOS) pour le japonais. Dans le tableau la formulation de XPOS est simplifiée.

2. Nous ne cherchons pas ici à savoir si il faudrait considérer les flexions pluriel (ou leur absence) comme des marques quantificationnelles, en particulier en l'absence de déterminant : *Adieu, veau, vache, cochon, couvée*

mot		SUD-POS	SUD-dep	XPOS	quantif
samazama-na	différents	ADJ	comp :aux	adjectif invariable	indéfini pluriel
kono,sono...	ce/ces/...	DET	det	démonstratif	défini
aru	un/des	DET	det	mot lié	indéfini
sorezore no	chacun des	NOUN	comp :obj	nom	distributif,défini
hitori no	un	NOUN	comp :obj	groupe numéral	indéfini
yama yama	montagnes	NOUN	_	nom commun	pluriel
kaku	chaque	NOUN	compound	préfixe	distributif
tachi	pluriel	NOUN	comp :obj	suffixe	pluriel, défini

TABLE 1 – Partie du discours des expressions quantifiantes (cas du japonais)

Les expressions quantifiantes sont présentes dans les strates lexicales natives et sino-CJ. Un déterminant peut apparaître dans les translitérations de noms propres «occidentaux» avec déterminant. Par exemple dans (ja) ラ . ファミユ *ra fami'yu* «La Famille», ザ・キング *za kingu* «The king». Ce sont des cas rares et il est permis de penser que sauf peut-être pour l'anglais, la valeur sémantique n'est pas connue des locuteurs ordinaires.

On note que certaines étiquettes des parties du discours, empruntées en grande partie à la notation UD (De Marneffe *et al.*, 2021), sont incohérentes avec les intitulés habituels. Par exemple (voir tableau 1), la catégorie NOUN assignée au quantificateur (co) *gak* / (ja) *kaku* heurte l'entendement commun puisque ce morphème n'a aucune des propriétés distributives et sémantiques des noms. On préférerait de loin l'étiquette «préfixe» de la grammaire usuelle. Celle-ci n'étant pas disponible dans la nomenclature en SUD, «DET» serait tout à fait satisfaisante. C'est d'ailleurs la partie de discours qui est attribuée à ce même quantificateur dans l'annotation disponible pour le chinois. [3]

3 Ajouter des informations relatives à la quantification

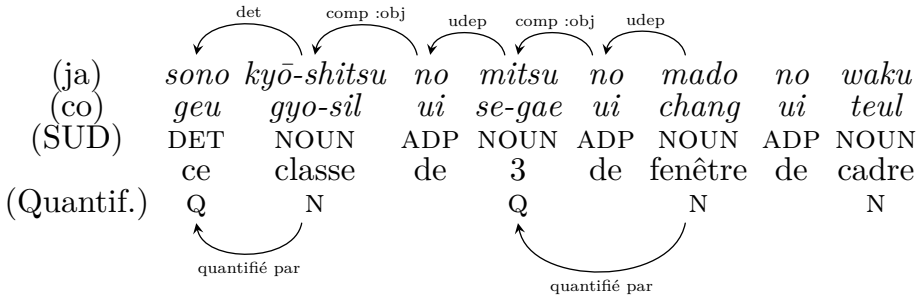
Comme indiqué en introduction, l'annotation relative à la quantification fournit trois informations : signaler les expressions nominales quantifiables, les expressions quantifiantes, et établir les liens de quantification entre les deux lorsqu'il y en a.

L'annotation en SUD ne comporte pas d'informations spécifiques sur la quantification. Mais plutôt que de modifier les données syntaxiques, nous préférons procéder à l'ajout de ces informations. L'utilisateur dispose alors des informations syntaxiques «natives» et des informations sémantiques additionnelles et peut disposer de l'une et l'autre à sa guise. Le principe est d'ajouter à chaque token une rubrique où est indiqué si le token est la tête (N) d'une expression nominale quantifiable, la tête (Q) d'une expression quantifiante ou autres (∅). Chaque token est doté d'une seconde rubrique additionnelle. Si le token est la tête d'une expression nominale quantifiée, cette rubrique contient un renvoi vers la tête de l'expression qui le quantifie. Si non, la rubrique vaut ∅. C'est le cas pour (co) *teul* / (ja) *waku* «cadre» dans l'exemple ci-dessous. [4]

3. (`zh-gsd-sud-train`) `各 各 DET DT _ 2 det _ LTranslit=gè|SpaceAfter=No|Translit=gè`

4. Dans l'exemple, l'analyse du groupe numéral avec compteur individuel ((co) 3_{NUM} gae_{CPTIND}) n'est pas celle adoptée dans le projet. Le traitement des compteurs individuels posant un problème qui dépasse le cadre de la présente présentation, nous adoptons ici une version simplifiée.

(2)



«le cadre des trois fenêtres de cette classe»

4 Problèmes de segmentation pour les lexiques sino-coréens et sino-japonais

Nous décrivons ici simplement la morphologie des lexiques sino-CJ, et les problèmes rencontrés avec la segmentation appliquée dans les treebanks aux formats UD et SUD. Nous proposerons dans la section suivante de modifier cette segmentation.

Le coréen et le japonais disposent de trois lexiques, chacun associé à un système morphologique qui lui est en partie propre : un lexique natif, un lexique d'origine chinoise et un lexique dit «étranger». Le lexique d'origine chinoise est basé sur un vocabulaire emprunté au chinois entre le Vè siècle et le XIXè siècle. Ce vocabulaire (phonologie, graphie, sémantique) a évolué avec le temps et a gagné en autonomie par rapport au chinois originel. Dans la langue contemporaine, ce vocabulaire est très présent tant pour la variété des formes que pour sa fréquence. Le vocabulaire étranger englobe tout le reste des emprunts, y compris ceux au chinois à partir du XIXè siècle.

Le système morphologique sino-CJ est abondamment décrit dans la littérature (voir entre autres pour une synthèse récente en japonais : Kobayashi *et al.* (2016)). Nous nous en tenons aux propriétés pertinentes pour la présente discussion. Pour simplifier, on peut dire que la «brique» de base du système morphologique sino-CJ est le morphème à une unité morpho-phonologique (KU). Ce morphème peut constituer à lui seul un mot ou être combiné pour constituer un mot sémantiquement transparent ou non. Par exemple (ja) 量 *ryō* «quantité» et (ja) 水 *sui* «eau» sont des KU. Le premier peut constituer à lui seul un nom commun. La combinaison des deux produit un nom commun sémantiquement transparent : *sui-ryō* «quantité d'eau ; litt. eau quantité». Par contre, (co) *sin-gyeong* / (ja) 神経 *shin-kei* «nerf» est constitué de deux KU (respectivement «dieux/esprit» et «passage/processus») mais n'est pas sémantiquement transparent.

Les KU sont de nature grammaticale variée. On parlera de KU à comportement nominal (KU_{NC}), verbal (KU_{V}), adjectival (KU_{ADJ}) ou quantificationnel (KU_{Q}) (voir Tab.2). Certains KU_{NC} sont faiblement autonomes : ils ne peuvent constituer un argument nominal mais possèdent plusieurs propriétés typiques du nom commun, parmi lesquelles celle d'être quantifiés. C'est le cas de (ja) 室 *shitsu* «salle» :

(3) 各室 [$kaku_{KU_Q}$ $shitsu_{KU_{NC}}$] «chaque salle»

A part les textes pour lecteur débutant, le texte brut japonais n'est pas segmenté. Il doit

Partie du discours	japonais	coréen	
KU_{ADJ}	新 *shin*	*sin*	nouveau
	新シリーズ *shin-shirīzu*	*sin sirijeu*	nouvelle série
KU_{V}	減 *gen*	*gam*	diminuer
	減量 *gen-ryō*	*gam ryang*	diminuer la quantité
KU_{NC}	量 *ryō*	*ryang*	quantité
KU_{Q}	各 *kaku*	*gak*	chaque
	各シリーズ *kaku-shiriizu*	*gak-sirijeu*	chaque série

TABLE 2 – Catégories grammaticales des morphèmes sino-japonais à une unité

l'être pour l'annotation automatique. La segmentation et l'annotation en parties du discours ont fait l'objet d'abondantes études. Cinq dispositifs sont en général cités (Omura *et al.*, 2021). Les treebanks en (S)UD sont proposés en deux versions, avec segmentation en unités dites «courtes» (SUW) et «longues» (LUW). On peut affirmer que la première est la plus utilisée.

Cependant, l'unité courte (et *a fortiori* l'unité longue) présente un défaut. Comme évoqué en introduction, les formes constituées de deux KU (ex : (ja) *kaku-shitsu* «chaque-salle») sont traitées comme des tokens, quand bien même elles sont compositionnelles et sémantiquement transparentes. Il faut donc adopter une segmentation plus fine sur ces formes composées.

Le texte coréen brut est segmenté. La segmentation utilisée en (S)UD reprend la segmentation conventionnelle commune. Avec cette segmentation, un nom et sa particule casuelle constituent un token. Park & Tyers (2019) a proposé une segmentation plus fine qui sépare les deux et se rapproche de la segmentation japonaise en SUW. Néanmoins, le traitement des formes sino-coréennes n'est pas uniforme Ainsi, les treebanks annotés en (S)UD contiennent des tokens constitués d'un KU_{Q} et d'un KU_{NC} (co) $[gakguk]_{\text{NOUN}}$ «chaque pays», $[gakjong]_{\text{NOUN}}$ «chaque espèce» etc. Des formes sémantiquement équivalentes peuvent au contraire être segmentées en deux tokens : $[gak_{\text{NOUN}}\ gun_{\text{NOUN}}]$ «chaque canton», $[gak_{\text{NOUN}}\ si_{\text{NOUN}}]$ «chaque ville». Une standardisation est souhaitable pour montrer les équivalences sémantiques entre les différentes formes, mais aussi les équivalences entre coréen et japonais.

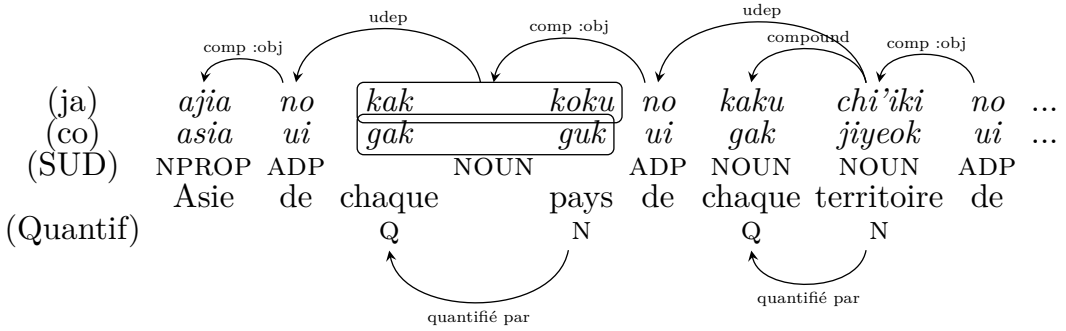
5 Resegmentation et annotation sémantique

Nous adoptons une segmentation similaire pour le japonais et le coréen, qui sépare les particules adnominales des noms. Nous avons choisi pour le japonais l'annotation en «unités courtes» et en coréen l'annotation MorphUD qui lui est très ressemblante. La similitude entre les deux langues devient alors plus frappante, au moins pour le sino-CJ.

Nous procédons comme suit. Dans un corpus segmenté et annoté en SUD, nous re-segmentons les tokens sino-CJ composés et sémantiquement transparents. Nous ajoutons les annotations concernant la quantification. Le schémas ci-dessous montre le résultat de cette opération. Les dépendances syntaxiques sont laissées en l'état. Nous proposerons en section 6 de les corriger. La segmentation nouvelle et l'annotation associée montrent que les mêmes relations sémantiques sont à l'oeuvre à l'intérieur du mot (ja : *kak-koku*) et à l'intérieur du groupe

nominal (ja : *kaku-chi'iki*).

(4)



«... de chaque territoire de chaque pays d'Asie»

Comme nous l'avons déjà dit, la segmentation des mots compositionnels n'est pas seulement intéressante lorsqu'il y a un quantifiant et un quantifié. Les mots composés d'un morphème nominal et d'un autre morphème nominal ou d'un morphème verbal méritent aussi d'être segmentés, en particulier pour contraster les langues avec ou sans quantificateur obligatoire. Par exemple, on observe dans le tableau Tab.3 que le sino-CJ ne quantifie pas les substantifs (KU_{NC}) à l'intérieur d'un nom [KU_{NC} KU_{NC}]$_{\text{NC}}$, mais qu'il arrive que le français le fasse dans les expressions sémantiquement équivalentes, au niveau syntaxique. On observe aussi qu'en français, cette quantification n'est pas systématiquement réalisée alors que les formes sont sémantiquement analogues. Dans cette distribution, la présence du quantificateur pourrait donc être une contrainte lexico-syntaxique sans motivation sémantique. En d'autres termes, nous serions en présence d'une forme de figements.

coréen	japonais	trad. littérale	français
[KU_{NC} KU_{NC}]			
su ryang	*sui ryō* 水量	eau quantité	quantité { ??de l'/de ∅} eau
su on	*sui on* 水温	eau température	température {de l'/* de ∅} eau
suu ap	*sui atsu* 水圧	eau pression	pression {de l'/* de ∅} eau
[KU_{V} KU_{NC}]			
jeung ap	*zou atsu* 増圧	aug. pression	augmentation de {la /?? ∅} pression

TABLE 3 – Comparaison sino-CJ et français de formes nominales composées

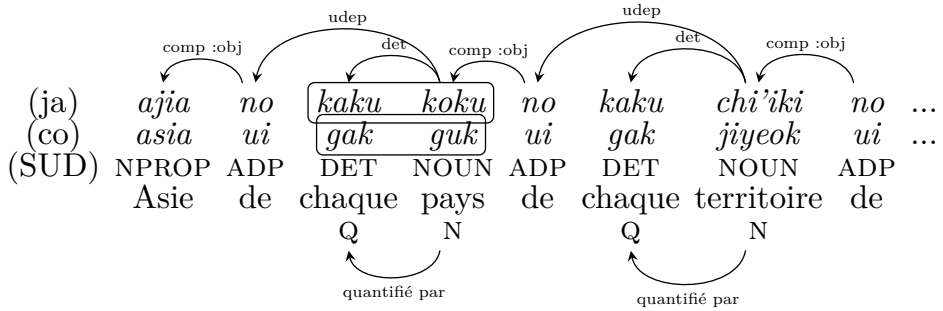
6 Correction des dépendances

En plus de l'intérêt pour l'analyse sémantique de la quantification, la segmentation que nous proposons profite aussi à l'analyse syntaxique, grâce à une plus grande précision.

Après segmentation du token, nous indiquons les dépendances entre les composants du token décomposé, et nous reportons les dépendances qui entraient et sortaient de ce token vers ses composants. Reprenons l'exemple (4). Les parties du discours sont ajoutées aux nouveaux composants. Nous étiquetons DET le quantificateur *kaku*, conformément à l'analyse

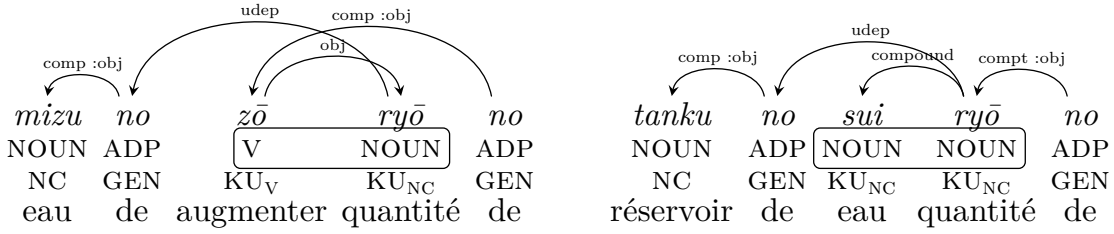
en linguistique et au chinois. Il serait désormais naturel de faire pointer la dépendance *udep* entrante sur le KU_{NC}. De même, la dépendance sortante *comp* : *obj* partirait de ce même morphème. Le schémas montre qu'en sino-CJ, les dépendances au niveau morphologique et syntaxique sont les mêmes et qu'elles correspondent à des relations sémantiques identiques.

(5)



Ces observations sont généralisables pour les autres formes avec au moins un KU_{NC} comme exemplifié ci-dessous (le coréen se comporte de la même manière, il n'est pas mentionné pour des raisons de place).

(6)



de l'augmentation de la quantité d'eau *de la quantité d'eau du réservoir*

7 Expérimentation

Nous avons évalué sur le japonais quels gains en finesse d'analyse étaient envisageables grâce à la nouvelle segmentation. La procédure est réalisée à l'aide de règles explicites. Elle est tout à fait opérationnelle pour le japonais mais inapplicable pour le coréen à cause de la présence massive de mots homophones, qui nécessite un traitement statistique.

Nous avons appliqué notre annotation sur deux corpus japonais (voir Tab.4). Le premier est un corpus pré-segmenté et annoté en SUD (`SUD_JapaneseGSD@2.15`, fichier `train`). Il contient ≈7000 phrases. Le second corpus est le sous corpus japonais du corpus comparable CompCor (Blin *et al.*, 2025). Il contient environ 23 millions de phrases provenant de Wikipedia et d'une collection d'articles de presse.

corpus	sous-corpus	Annotation	# phrases ja	
`SUD_JapaneseGSD`	Wikipedia	Dépendance, format SUD	7 050	
CompCor	Wikipedia	texte brut	$\approx 21M$	$\approx 6M$
	Article de presse	texte brut	$\approx 2M$	$\approx 600K$

TABLE 4 – Description des corpus

La procédure est gérée par des règles (implémentées en Python). Elle suppose une pré-segmentation et annotation en parties du discours XPOS, conforme à celles d'UniDic. L'information est déjà disponible dans le corpus GSD. Pour ce qui est du texte brut du CompCor, il est préalablement segmenté et annoté à l'aide de MeCab (Kudo, 2006) [5] associé au dictionnaire UniDic [6] . Il faut par ailleurs disposer de la liste des KU avec au moins leur graphie et leur lecture. Nous ne disposons que d'une liste partielle des KU_{NC} car paradoxalement, bien que très présents dans la langue, ils sont encore peu étudiés. Une liste est en cours d'élaboration et comprend actuellement environ 300 morphèmes, qui peuvent être conçus comme des triplets «graphie-prononciation-sens». Étant donné la présence d'homographes et/ou homophones, la liste se réduit à 200 paires «graphe-lecture» exploitables pour un traitement automatique. Les KU_{Q} constituent une liste petite et fermée. Nous nous sommes par ailleurs donné des listes de quelques KU_{ADJ} et KU_{V} dont la capacité à composer un mot avec un KU_{NC} est avérée. Nous ne risquons ainsi pas de sur-décomposer.

Une fois le texte segmenté, il faut parcourir la liste des tokens de la phrase, et pour chacun vérifier si il est composé de deux KU, dont un KU_{NC}. Si oui, déterminer si l'autre composant est un KU_{V}, un KU_{ADJ} ou encore un KU_{NC}. En déduire la relation de dépendance (voir exemples 6). Pour une forme [KU_{NC} KU_{NC}], il est nécessaire de connaître la nature sémantique des composants et leur valence, nulle ou non. Si un des KU_{NC} a une valence non nulle en déduire que l'autre est son complément. Autrement, il s'agit d'une coordination (ex. : *dan-jo* «homme et femme», *to-dō-hu-ken* «capitale, Hokkaïdō, département et région».

En appliquant cette procédure, nous constatons que entre 4% et 5,5% des phrases contenues dans les corpus utilisés contiennent un nom décomposable et pourraient être corrigées.

	type de nom			phrases
	[KU_{Q} KU_{NC}]	[KU_{ADJ} KU_{NC}]	[KU_{V} KU_{NC}]	modifiées
SUD_JapaneseGSD	169	26	123	4,2%
CompCor	73 935	32 252	52 708	5,4%

TABLE 5 – Nombre d'occurrences de noms décomposables et gains observés en japonais.

8 Conclusion

Nous avons présenté la première étape du projet de constitution d'un corpus comparable, annoté pour l'étude sémantique de la quantification en chinois, coréen, français et japonais. Il s'agissait de choisir un mode de segmentation pour le coréen et le japonais et d'observer ses effets sur la finesse d'analyse. Nous avions en effet constaté qu'il était nécessaire d'adopter une segmentation plus fine que la segmentation en général utilisée dans les treebanks au format (S)UD. Pour le japonais, cette re-segmentation est facile à réaliser automatiquement par règles à partir de la segmentation existante. Pour le coréen, nous avons échoué avec une analyse à base de simples règles. Il est nécessaire manifestement de recourrir à des modèles statistiques. La nouvelle annotation que nous proposons permet de gagner en finesse d'analyse et fait mieux apparaître les régularités à l'intérieur de chaque langue et d'une langue à l'autre.

5. https://github.com/taku910/mecab
6. https://clrd.ninjal.ac.jp/unidic/download.html

Le travail était volontairement concentré sur traitement du vocabulaire sino-CJ. La question de la segmentation semble moins problématique pour les autres strates lexicales. Un point à régler est l'annotation des redoublements de mots à valeur pluriel, comme par exemple (ja) *yama-yama* «les montagnes ; litt. montagne/s-montagne/s». Nous devons ensuite évaluer la pertinence de la segmentation en chinois.

Références

Blin R., Delaporte A., Wang I., Arslangul A., Yu X. C. & Noûs C. (2025). CompCor-v0.1.2 : Corpus Comparable Coréen Français Japonais Mandarin. Description du corpus CompCor-v0.1.2, hal : hal-04864542.

Bond F. (2001). *Determiners and Number in English contrasted with Japanese, as exemplified in Machine Translation*. Thèse de doctorat, University of Queensland.

Chierchia G. (1998). Reference to kinds across language. *Natural language semantics*, (4), 339–405. doi : 10.1023/A :1008324218506.

De Marneffe M.-C., Manning C. D., Nivre J. & Zeman D. (2021). Universal dependencies. *Computational linguistics*, **47**(2), 255–308.

Den Y., Ogiso T., Ogura H., Yamada A., Minematsu N., Uchimoto K. & Koiso H. (2007). The development of an electronic dictionary for morphological analysis and its application to japanese corpus linguistics [in japanese]. *Nihongo Kagaku*, p. 101–123.

Gerdes K., Guillaume B., Kahane S. & Perrier G. (2019). Improving surface-syntactic universal dependencies (sud) : surface-syntactic relations and deep syntactic features. In *TLT 2019-18th International Workshop on Treebanks and Linguistic Theories*, p. 126–132 : Association for Computational Linguistics.

Kobayashi H., Yamashita K. & Kageyama T. (2016). Sino-japanese words. In *Handbook of Japanese Lexicon and Word Formation*, p. 93–131. De Gruyter Mouton, kageyama ,taro and kishimoto,hideki édition. doi : https ://doi.org/10.1515/9781614512097.

Kudo T. (2006). Mecab : yet another part-of-speech and morphological analyzer.

Kuno S., Takami K.-i. & Wu Y. (1999). Quantifier scope in english, chinese, and japanese. *Language*, (1), 63–111.

Marsden H. (2009). Distributive quantifier scope in english-japanese and korean-japanese interlanguage. *Language Acquisition*, **16**(3), 135-177. doi : 10.1080/10489220902967135.

Omura M., Wakasa A. & Asahara M. (2021). Word delimitation issues in ud japanese. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, p. 142–150.

Park J. & Tyers F. (2019). A new annotation scheme for the Sejong part-of-speech tagged corpus. In A. Friedrich, D. Zeyrek & J. Hoek, Éds., *Proceedings of the 13th Linguistic Annotation Workshop*, p. 195–202, Florence, Italy : Association for Computational Linguistics. doi : 10.18653/v1/W19-4022.