CompCor-v0.1.2 : Corpus Comparable Coréen Français Japonais Mandarin

Raoul Blin, Alexander Delaporte, Ilaine Wang, Arnaud Arslangul, Xinyue Cécilia Yu, Camille Noûs

CompCor-v0.1.2 : Corpus Comparable Coréen Français Japonais Mandarin

Raoul Blin, Alexander Delaporte, Ilaine Wang, Arnaud Arslangul, Xinyue Cécilia Yu, Camille Noûs
{blin,alexander.delaporte}@ehess.fr
{ilaine.wang,arnaud.arslangul,xinyue.yu}@inalco.fr

Table des matières

1 Introduction

Le présent document décrit le corpus `CompCor`, version `0.1.2`. Le `CompCor` («Corpus Comparable coréen, français, japonais, mandarin») rassemble des corpus comparables monolingues natifs en quatre langues : coréen, français, japonais et mandarin chinois. Les corpus sont comparables du point de vue du type de production. La version `0.1.2` comprend un corpus journalistique et l'encyclopédie Wikipédia. Les tailles des sous-corpus varient en fonction de la disponibilité des ressources. Le corpus est en accès partiellement libre, en fonction de l'origine des textes.

Le présent document fournit un argumentaire, un descritif du contenu et du format, et des informations sur la diffusion du corpus.

2 Argument du projet

Le projet `CompCor` vise à mettre à disposition des linguistes un corpus multilingue, comparable, de bonne qualité. Quatre langues sont concernées : deux langues d'Asie orientale, de type SOV (coréen, japonais), une troisième langue d'Asie orientale mais de type SVO (mandarin continental) et une langue européenne SVO (français).

L'objectif étant de servir de ressources à des études en linguistique, le corpus doit obéir à plusieurs contraintes

(1) Les données sont nativement produites dans les langues étudiées afin de ne pas introduire des biais de traduction : textes peu naturels influencés par la langue source par exemple.

(2) Les données sont en quantité suffisante pour donner une image aussi juste que possible de chacune des langues. La volume important réduit la visibilité des hapax, des cas «limites» peu représentatifs ou des erreurs. La quantité augmente aussi la significativité des résultats obtenus par des études quantitatives

(3) Les données appartiennent à des styles et registres de langues comparables et limite les biais de style.

(4) Les données sont formatées de façon uniforme pour faciliter les comparaison et les traitements automatiques.

Aujourd'hui, une telle ressource n'existe pas, *a fortiori* pour les langues choisies dans le présent projet. Il existe bien des corpus «traduits», numérisés et librement disponibles [1]. Mais la plupart sont synthétiques, traduits à partir d'un corpus monolingue, en général l'anglais. C'est le cas notamment des corpus multilingues des conférences TED (voir par exemple Reimers and Gurevych [2020]). L'influence de la langue source sur la langue cible n'est pas discutée. De surcroît les traductions de bonne qualité sont peu nombreuses et rarement (voire jamais à notre connaissance) évaluées par des traducteurs professionnels. Les corpus traduits de grande taille sont en général extrêmement bruités (Blin [2018]) et en grande partie inexploitables, pour quelque usage que ce soit.

Enfin, les ressources traduites sont en générale alignées phrase par phrase. Les formats utilisés ne préservent pas les informations relatives à la structure du texte dont elles sont issues. C'est une absence qui peut pénaliser des études nécessitant des observations sur des segments de plusieurs phrases, comme les études d'anaphores par exemple. Il est donc nécessaire de disposer de textes non segmentés (par phrases) et mettant à disposition des structures textuelles complètes (paragraphes, sections etc).

Le choix s'est porté sur un ensemble de langues individuellement bien dotées en ressources numérisées monolingues, mais très inégalement comparées entre elles dans les travaux de linguistique contemporaine. Les paires japonais-chinois (mandarin compris) et peut-être plus encore japonais-coréen font l'objet de nombreux travaux comparatifs. Par exemple un ouvrage récurrent (Japanese/Korean Linguistics, Ed. CSLI publications) est consacré au japonais-coréen. Les paires impliquant ces langues et le français sont moins étudiées. Les publications sont sporadiques et plutôt limitées à des publics spécialisés sur ces langues. Enfin, il n'existe pas à notre connaissance de travaux impliquant les quatre langues. En général, les travaux comparant les langues orientales et occidentales s'en tiennent à l'anglais et au mandarin, d'ailleurs souvent implicitement tenus comme des archétypes. Typiquement Chierchia [1998] par exemple regroupe mandarin, coréen et japonais sous une même étiquette «langues à classificateur» et prend le mandarin comme exemple, oubliant de mentionner les différences morphosyntaxiques notoires entre ces langues, et oubliant d'évaluer les effets de ces différences sur sa théorie. Ce type d'étude pourrait grandement profiter de l'existence de corpus comparables, au moins pour légitimer le (non)rapprochement entre langues.

3 Caractériser les corpus

Nous proposons ici quelques traits permettant de caractériser et comparer les types de corpus, et de façon pertinente pour des études linguistiques. Nous reprendrons ces traits dans les sections suivantes, où sont présentés les sous-corpus. Tous les traits ne sont pas objectivement évaluables.

1. Pour se faire une idée, consulter par exemple la page `opus.nlpl.eu`, Tiedemann and Nygaard [2004]

Traits	valeurs possibles	
Niveau de correction	spontané	aucune correction
	autocorrection	corrigé par l'auteur ; style pouvant donc rester «personnel»
	correction collective	corrigé par autrui, le style peut-être moins personnel, obéissant certainement à des normes[(a)]
Média[(b)]	écrit	
	oral verbatim	retranscription brute (ex : CHILDES (MacWhinney [1992]))
	oral textualisé	discours rapporté
Interaction	récit	sans adresse à un interlocuteur
	monologue	à l'adresse d'un public bien identifié mais sans perspective d'interaction verbale ; ex : conférences
	dialogue	conversation libre

TABLE 2 – Proposition de traits pour comparer les et classer les corpus.

(a) La «correction collective», correction par des tiers, implique certainement une norme. Par exemple, le style journalistique obéit à certaines normes qui ne sont pas celles de la rédaction de brevets.
(b) On peut s'interroger sur l'existence d'un format intermédiaire entre l'oral et l'écrit, celui des micro-blogging (SMS, *chat*,etc.).

4 Corpus

La version `0.1.2` du `CompCor` comprend deux types de textes : corpus journalistique et encyclopédique.

4.1 Articles de journaux

Traits : écrit, correction collective, récit

C'est un corpus incontournable car présent dans toutes les langues étudiées. Leur mode de production et leur forme sont très proches d'une langue à l'autre.

Ils sont produits par un auteur unique en général, ou un petit collectif d'auteurs (deux ou trois maximum), et sont en général relus par des tiers et éventuellement corrigés. Les journaux peuvent imposer un style (contraintes lexicales et morphosyntaxiques). Ce sera donc une convention collective et non pas individuelle.

Structurellement, les articles ont une structure similaire d'une langue à l'autre : ils sont constitués au plus d'un titre, un sous-titre, un résumé et un corps de texte.

Le corpus rassemble les versions en ligne de journaux de diffusion nationale (*vs* régionale etc.). Sont retenus les journaux dont la version papier connaît le plus fort tirage. La collecte a par ailleurs été limitée aux journaux dont la consultation automatique est autorisée (consultation du fichier `robot.txt`). Aucun ne peut cependant être reproduit.

langue	Titre	Rediffusable	Source	genre
Coréen	Chosun Ilbo	non	chosun.com	grand public
	Dong-A Ilbo	non	donga.com	grand public
	The Joong Ang	non	joongang.co.kr	grand public
Français	Libération	non	liberation.fr	grand public
	Les échos	non	lesechos.fr	économique
	Le figaro	non	www.lefigaro.fr	grand public
	L'opinion	non	www.lopinion.fr	grand public
	La croix	non	www.la-croix.com	grand public
Japonais	Mainichi	non	mainichi.com	grand public
	Asahi	non	asahi.com	grand public
	Nikkei	non	nikkei.com	économique
	Sankei	non	sankei.com	grand public
Mandarin	Quotidien du peuple	non	XXX.people.com.cn	
	xinhuanet	non	www.xinhuanet.com	
	ce.cn	non	www.ce.cn	
		non	www.rmxiongan.com	

4.1.1 Format des données

Le formatage xml des articles préserve la structure en minimisant les annotations. Sont distingués le titre (titre et sous-titre), un possible résumé, et le corps de l'article.

Quelques contenus sont modifiés pour prévenir les biais de comptages. Les articles peuvent être tronqués, en particulier lorsque l'accès est payant. Dans ce cas, nous insérons une balise `<texte_tronqué/>`. Si la troncation se fait au milieu d'un paragraphe, le paragraphe est éliminé. Toute information éditoriale, susceptible d'apparaître dans plusieurs articles, est extraite du texte et mise à part. C'est le cas du nom des auteurs et d'autres informations relatives à l'édition. Par ailleurs, certains titres contiennent des noms de rubrique, que l'on retrouve d'un article à l'autre. Ces noms de rubriques sont isolés et passés en valeur d'un attribut `metadata` ou `cat` [2] . Par exemple,

丰都：引名校办学促进教育高质量发展

deviens

```
<title metadata=" 丰都 "> 引名校办学促进教育高质量发展 </title>
```

4.1.2 Commentaires sur la ressource

Pour le français et le japonais, les articles rassemblés sont ceux des journaux dont les versions papier sont les plus diffusées d'après les chiffres officiels (disponibles sur Wikipédia notamment) et dont la consultation automatique en ligne est autorisée. On peut donc dire que le corpus est représentatif des journaux officiellement les plus diffusés. Le corpus mandarin est très largement dominé par les versions nationales et locales du Quotidien du Peuple.

4.2 Wikipedia

Traits : écrit, correction collective, récit

L'encyclopédie en ligne est une ressource incontournable pour tout projet de corpus. Elle est de grande taille, de bonne qualité et librement accessible (licence libre). On a utilisé les fichiers : `XXwiki-20240920-pages-articles-multistream.xml`.

4.2.1 Format des données

La structure des pages est maintenue. Dans le corps de texte, seul le texte et les titres de section sont conservés : les tableaux, figures et autres sont éliminées et leur emplacement est marqué par des points de suspensions.

2. Les deux balises doivent être confondues à terme.

4.2.2 Commentaires sur la ressource

Il s'agit d'un type de texte très particulier. Le niveau de correction varie certainement d'un texte à l'autre.

On peut s'attendre à rencontrer un vocabulaire globalement plus riche qu'ailleurs puisque par définition, une encyclopédie aborde tous les sujets.

En mandarin, on observe que les caractères simplifiés coexistent avec les caractères traditionnels. Pour harmoniser le corpus, n'ont été finalement retenus que les textes ne contenant que des caractères simplifiés [3]

Après traitement, les textes contiennent encore de nombreuses erreurs.

5 Récapitulatif

Le nombre de documents par corpus est réparti comme suit :

	ko	fr	ja	zh	total
Presse	23 972	223 483	172 877	234 760	655 092
Wikipedia	1 176 172	3 886 518	1 858 068	233 560	7 154 318
Total par langue	1 200 144	4 110 001	2 030 945	468 320	7 809 410

TABLE 3 – # pages par corpus

	ko	fr	ja	zh	total
presse	621 364	5 167 493	1 917 060	3 699 920	11 405 837
wikipedia	7 196 921	34 720 843	21 086 798	763 331	63 767 893
totalparlg	7 818 285	39 888 336	23 003 858	4 463 251	75 173 730

TABLE 4 – # phrases par corpus

6 Encodage des caractères

Les corpus sont tous encodés en `utf8`. Mais à l'intérieur de cet encodage, certains caractères peuvent être encodés sur 1, 2 ou trois octets. C'est par exemple le cas des caractères latins et des numéraux en japonais, que l'on trouve sous la forme «`1`» (un octet) ou «　`1`　» (trois octets). D'un point de vue informatique, deux mêmes graphèmes encodés différemment sont deux objets distincts. Les analyseurs peuvent les traiter différemment et en général, l'encodage est unifié avant le traitement d'un corpus. Mais pour le `CompCor`, faute de disposer d'une vue suffisamment claire pour toutes les langues, les encodages ont été laissés en l'état.

7 Versions

Le corpus est amené à évoluer dans le temps. En conséquence, le corpus est versionné.

Les versions sont encodées avec trois chiffres : X.Y.Z . X est incrémenté de 1 pour un ajout de style ou de langue. Y est incrémenté de 1 pour un ajout de texte à l'intérieur d'une langue ou d'un style existant. Z est incrémenté à chaque modification de données existantes.

Le numéro de version suffixé de **public** si la version est publique.

v0.1.2 , 2025/01/27

Corrections mineures dans le manuel.

v0.1.0 , 2024/11/20

Première version (β).

3. Pour cela, on a transcrit à l'aide de `opencc` les textes en caractères simplifiés. Si un texte ainsi converti est différent de l'original, c'est qu'il contient originellement des caractères non simplifiés. Il est alors exclu.

8 Licence

Une partie seulement du `CompCor` est librement diffusée. Pour la partie libre, la licence est dans la mesure du possible du type Creative Commons. Par défaut, les textes sont livrés dans les conditions de leur licence d'origine.

9 Disponibilité et téléchargement

Les corpus sont disponibles aux adresses ci-dessous. Seule la dernière version publique, est accessible. Pour accéder aux versions antérieures ou non publiques du corpus, merci de contacter `blin@ehess.fr`.

— `v0.1.2 (public)` : https://sharedocs.huma-num.fr/wl/?id=JrUaXBoJYig2xiiKAtTDYYWZSzIVNOR3

— `Manuel, v0.1.2` : https://sharedocs.huma-num.fr/wl/?id=nSDpu1YxoZQZoo3dgaslh9qsxY5O2YoA

10 Liste de travaux basés sur le `CompCor`

coming soon !

11 Contributions

Les personnes suivantes ont participé au projet. R.Blin coordonne le projet et est en charge plus particulièrement de la compilation des données japonaises et mandarin. A.Delaporte est en charge des données françaises et I.Wang des données coréennes. Les discussions sur le projet, sur le format et le choix des données, incluent aussi A.Arlangul et Xinyue Yu (notamment pour l'expertise sur le mandarin).

Le projet bénéficie aussi des contributions des participants au séminaire de l'axe 2 du Centre de Recherches sur les Langues de l'Asie Orientale (CRLAO).

Références

Raoul Blin. Automatic evaluation of alignments without using a gold-corpus - example with french-japanese aligned corpora. In Shirai Kiyoaki, editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-24-5.

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

Brian MacWhinney. The childes project : tools for analyzing talk. *Child Language Teaching and Therapy*, 8(2) : 217–218, June 1992. ISSN 1477-0865. doi : 10.1177/026565909200800211. URL http://dx.doi.org/10.1177/026565909200800211.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. URL https://arxiv.org/abs/2004.09813.

Jörg Tiedemann and Lars Nygaard. The OPUS corpus - parallel & free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004. URL http://stp.ling.uu.se/~joerg/paper/opus_lrec04.pdf.