Building and Annotating a Large Comparable Corpus for Studying Semantic Quantification - Chinese, French, Japanese, Korean

Raoul Blin[(1)], **Jinnam Choi**[(2)], **Qishen Wu**[(3)], **Yuxin Zhang**[(4)], **Soonhee Hwang**[(5)], **Takahiro Morita**[(6)], **Alexander Delaporte**[(1)], **Ilaine Wang**[(7)], **Chang Liu**[(8)]
[(1)]CNRS-CRLAO, Aubervilliers, [(2)]CLLE, Université Jean Jaurès,
[(3)]Université Paris Nanterre, [(4)]STIH, Sorbonne Université, [(5)]Hongik University,
[(6)]Kyoto University, [(7)]ERTIM-INALCO, [(8)]INALCO
[(1)]Aubervilliers, France, [(2)]Toulouse, France, [(3)]Nanterre, France,
[(5)]Sejong, Korea, [(6)]Kyoto, Japan, [(4,7,8)]Paris, France
{blin, alexander.delaporte}@ehess.fr, {jinnamchoi11, yuxinab.zhang}@gmail.com,
qishen.wu@sorbonne-nouvelle.fr, soonheehwang@hongik.ac.kr,
morita.takahiro.8r@kyoto-u.ac.jp, {ilaine.wang, chang.liu}@inalco.fr

Abstract

Quantifiers and noun quantification are well-studied topics in linguistics, but, to the best of our knowledge, there are still no dedicated multilingual resources for the study of quantification. To address this gap, we compiled a large multilingual comparable corpus (Chinese, French, Japanese, Korean) and propose to enrich it with both syntactic and "quantificational annotation" (semantic information relevant to the study of quantification). In this paper, we present both the corpus and the annotation project, and report on our initial attempt at quantificational annotation, the challenges encountered, and the linguistic observations drawn from it.

Keywords: comparative corpus, quantificational annotation, Chinese-French-Japanese-Korean

1. Introduction

The expression of quantification (definiteness, quantity, or distributivity) in common nouns has been extensively studied in linguistics. This topic has given rise to influential publications across several subfields, including syntactic theory (e.g., Abney, 1987) and semantics (e.g., Barwise and Cooper, 1981, among others). The applied domain is no exception, with studies in machine translation (Bond, 2001) and language acquisition (Cho and Slabakova, 2014, etc.).

Despite the keen interest in the topic, there are no existing resources designed for large-scale study of quantification —whether from a syntactic or a semantic perspective. Even more so, there are no such resources for comparing quantification strategies across languages. Existing comparative studies typically rely on a small set of languages, often including English, and are heavily influenced by theories developed from observations of this specific language. For example, only determiners are analyzed, since they are the primary means of expressing quantification in English. Such studies often overlook the fact that in other languages, quantification may be expressed through different morpho-lexico-syntactic categories. In philosophical semantics, a widely cited work by Chierchia (1998) illustrates this approach. The author proposes a "typology of kind reference", focusing on determiner interpretation and the distinction between languages with obligatory determiners and those without. His typology covers many languages, including the ones that interest us here: Chinese, Japanese, Korean, and French.

However, the data provided are extremely limited: examples are few and quite simplistic. The data are also heavily simplified. For instance, Chinese is taken as representative of so-called "classifier languages", a group that supposedly includes Japanese and Korean. Yet Chinese, on the one hand, and Japanese and Korean, on the other, are linguistically distant. Grouping them together would require a more thoroughly argued and documented justification. Without questioning the hypotheses of this author and those who build on his work, it is fair to regret that such claims are not supported by large-scale quantitative and qualitative evidence.

We started a project aimed at addressing the lack of suitable resources for comparative studies of nominal quantification. The goal is to build a large, comparable multilingual corpus enriched with "quantificational annotation" —that is, annotation that identifies at least the quantifying expressions ("quantifiers"), the quantifiable nominal expressions ("QNE"), the quantified QNEs, the logical links between them (which quantifier quantify which QNE), and coreference chains across discourse.

In this paper, we present the first step of the project: the compilation of a raw, comparable corpus and the partial quantificational annotation of

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6685–6694
11–16 May 2026.

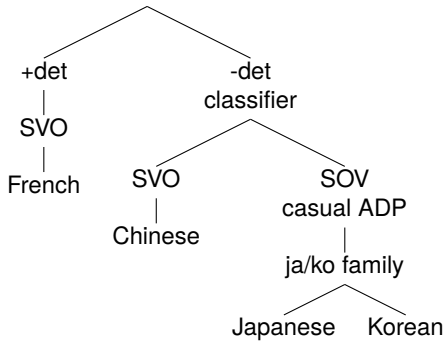


Figure 1: Distribution of the languages according to their main properties; "+/- det" indicates (quasi-)obligatory or non-obligatory determiners; "ADP" stands for "adpositions".

a reference subcorpus drawn from this raw data. This partial annotation includes identifying quantifying and quantified expressions, and linking them.

The paper is structured as follows. In Section 2, we briefly review the morphosyntactic features of the languages under study (Japanese, Korean, Chinese, and French). Section 3 outlines the specifications that guided corpus selection, followed by existing resources (Section 4) and the final corpus adopted (Section 5). Section 6 describes the annotation format. We then present (Section 7.2) the main principles of quantificational annotation. For detailed guidelines, we refer readers to the annotation guideline [1]. To assess the feasibility of such annotation, we applied it to a sample set. Section 7 describes this process. This experiment resulted in a reference corpus that can serve either as an evaluation corpus or a seed set for bootstrapping language models. It also provided quantitative data supporting typological hypotheses such as Chierchia's (see above). We report the annotation statistics in Section 7.3. Finally, the annotation experiment revealed a number of analytical challenges involved in quantificational annotation. These are discussed in Section 7.4.

2. Overview of the Languages Under Study

The project focuses on four languages: Chinese, French, Japanese and Korean. This selection allows for the study of quantification in both typologically distant languages —including a non-English Indo-European language —and in closely related ones. (see Fig.1).

These languages can be divided into several subgroups, as shown in Fig.1. A key distinction lies between languages with a syntactically (quasi-)obligatory determiner in French and those where determiners are not obligatory in Chinese, Japanese, Korean. This property is notably correlated by (Chierchia, 1998) with the presence of classifiers —absent in the former group and quasi-obligatory in the latter. Another grouping criterion is word order: French and Chinese follow an SVO order, while Japanese and Korean are SOV languages. This syntactic feature is in turn correlated with the adposition: absent in the former group, and systematically present in the latter.

Chinese, Japanese, and Korean (hereafter abbreviated as CJK) share a partially common vocabulary and associated morphological system. This shared vocabulary originated from extensive borrowing from Chinese up until roughly the early 19th century. Since then, it has evolved relatively independently within each language, although cross-borrowings have occurred later on —particularly in the late 19th and early 20th centuries. During that period, for instance, many Sino-Japanese terms were (re)introduced into Chinese. Sino-Japanese and Sino-Korean vocabulary remain highly prevalent and productive in their respective languages today.

The category of QNEs includes several subcategories: native common nouns, verbal nouns, and monomorphemic nominals. Verbal nouns have the dual ability to function either as common nouns or as verbs. In Japanese and Korean, verbal usage is marked by the presence of a light verb. In Chinese, it is inferred either from the syntactic position (see example below) or from the presence of a experiential aspect marker, such as *guò*.

(1) (ch) yīnwèi xǐhuān xuéxí suǒyǐ xuéxí
because love study then to study
As I love studying (litt. studies) I study.

Verbal nouns —especially those of Chinese origin —can be concatenated with other nouns to form a noun phrase. In this distribution, there is no formal marker that clearly indicates whether the expression is being used verbally or nominally. We treat such cases as nominal:

(2) (ko) somang pyohyeon "litt. to/wish - to express/expression; expression of a wish"

"Nominal monomorphemes" are single-phonological-unit morphemes that carry meaning. In Chinese and Sino-Japanese, they are typically written using a single character —for example, (ja) 量 *ryō* "quantity". Some monomorphemes behave like bound morphemes, functioning similarly to suffixes. Nevertheless, they are fully quantifiable like regular nouns. For instance, the Japanese morpheme *sha* "car" cannot appear on its own, but it can occupy a nominal argument

[1] The guideline and the annotated sample are available at https://sharedocs.huma-num.fr/wl/?id=vCJs5wTrUKFV5EjZPwLAiRqnu3tiF4LS

position when quantified by a distributive prefix: *kaku-sha* “each car”. These monomorphemes can also be combined to form compound words, such as (ja) *sui-ryō* “water-quantity; amount of water”.

As for quantifying expressions, we consider any expression that explicitly encodes definiteness and/or number and/or distributivity. In French, this mainly concerns determiners, including nominal groups used as quantifiers (see Ex. 4). In CJK languages, this also includes prefixes —such as (ja,ch) 各, (ko) *gak* —and suffixes (ja) *tachi*, (ko) *deul*.

3. Corpus design specifications

Building a corpus for the study of quantification involves two stages: selecting a “raw” corpus and later adding quantificational annotations (Section 6).

To support cross-linguistic comparison of the syntax and semantics of quantification, the corpus must meet the following criteria:

1. Authenticity : The corpus must consist of attested, natural language data; synthetic or artificially generated texts are excluded.

2. Comparability : For cross-linguistic comparison, corpus must represent comparable sublanguages, produced under similar conditions and for similar audiences.

3. Representativeness : The corpus should accurately reflect the genres it covers. Whenever possible, an exhaustive collection of a genre is preferable to a sampled one, since determining a representative sample size is difficult. Modern tools now allow for management of complete collections and the assessment of annotator reliability (annotator’ s agreement), facilitating large-scale data handling and consistency.

4. Usability and Format : Data should be provided in a minimally processed, reversible format that preserves the original structure and segmentation of texts, ensuring usability across theoretical frameworks.

5. Context Preservation : Since quantificational interpretation (e.g., definiteness) depends on cotext and discourse position, full texts and their organization (chapters, paragraphs, speaker turns, titles, etc.) must be retained.

6. Discourse Structure : The position of an expression within the discourse (e.g., in a title, introduction, or conclusion) affects its quantificational value. The corpus must therefore preserve discourse organization and indicate speaker turns in transcribed dialogues.

7. Open Access : The resource should be freely available and distributable.

4. Existing resources

Currently, for the four languages, with the exception of CompCor (see next section), no resources meet the specifications outlined in the previous section. The only existing corpora are aligned corpora [2] which fail to meet almost all of the requirements.

Indeed, these are synthetic, translated corpora. Large-scale ones (with over a million sentences) are typically extremely noisy and largely unusable for any serious application (Blin, 2018). Corpora of acceptable quality (intuitively judged, as no formal evaluations are provided) tend to be small and do not offer sufficient data. They are mostly translated from English. This is the case, for example, with the TED corpus (Reimers and Gurevych, 2020), which is available in multiple languages. Another example of a high-quality translated corpus is the ASPEC scientific article corpus (Nakazawa et al., 2016). However, it has two major drawbacks: it covers only Japanese, Chinese, and English, and it does not provide full texts.

Large Language Models (LLMs) now make it possible to generate parallel texts of high quality and in unlimited quantity. However, this raises three key issues. First and foremost, the goal of this project is to provide a resource for the analysis of natural language, not machine-generated language. Second, the translations would need to be evaluated, which is not a trivial task. Given the large size of the corpus, a full manual evaluation is materially impossible. Moreover, there are no automatic evaluation tools that are entirely satisfactory. For these reasons, LLM outputs cannot be used as primary resources. This does not preclude their use for generating ad hoc, targeted examples that can be validated by a human, especially when such examples are absent from the corpus.

As for comparable corpora composed of “natural” texts, resources already used in research are almost nonexistent, and none covers all four languages. To our knowledge, the only comparable corpus involving French and an Asian language is the one described in Raithel et al. (2024), which includes only Japanese and French. Its size is very modest (about 4,200 tokens in French and 3,000 characters in Japanese).

In both French and Japanese at least, there are digitized literary corpora that are in the public domain[3] . These could be brought together to form a comparable corpus, but the challenge is twofold. First, the category of “literary genre” is

[2]For an overview, see for example the page opus.nlpl.eu, (Tiedemann and Nygaard, 2004)

[3]For example, in French: abu.cnam.fr; in Japanese: www.aozora.gr.jp

extremely broad and involves highly diverse forms of language (styles, time periods, morphosyntax, vocabulary, etc.), making it difficult to build a homogeneous corpus. It would be necessary to group texts by genres that are common across all the languages. To our knowledge, such sorting has never been applied to the literature of the four languages we are focusing on. Second, due to legal copyright restrictions, only texts that are at least fifty years old are available. This poses a problem, as some languages have evolved significantly. This is particularly true in the case of Japanese. Most of the authors featured in the remarkable Aozora collection were active before the Second World War. They use a language typical of that era, which is not representative of contemporary usage.

5. CompCor corpus

To our knowledge, only the CompCor [4]corpus (Blin et al., 2025) meets the specifications outlined in the previous section. We present its characteristics below, following the structure of the criteria.

1) The corpus consists of texts in "human" natural language. There is one exception worth mentioning: CompCor includes the translation of the first two chapters of *The Little Prince*. The translations are from French into the CJK languages, carried out by native speakers of the target language, and is of professional quality. *The Little Prince* represents a negligible portion of the overall corpus and does not meaningfully affect the corpus wide textometric statistics. It can simply be excluded from analyses if necessary.

The interest in *The Little Prince* lies in the fact that it has been translated into a large number of languages and is used in several corpus linguistics studies, including those involving languages that are rarely compared to French. Among these languages are Chinese (see, e.g., Peng et al., 2020) and Korean (e.g., Hwang et al., 2020).

2) CompCor is comparable, not parallel. Only *The Little Prince* subcorpus is parallel and aligned paragraph by paragraph. However, as noted above, this subcorpus is very small and has a negligible influence on the results of experiments conducted on the corpus as a whole.

3) The corpus is not representative of each language in its entirety. It is not balanced, as it contains only written language. Moreover, it only includes two main genres: encyclopedic and journalistic texts; *The Little Prince* being very small in comparison as mentioned. For every language except Chinese, the corpus is dominated by Wikipedia, which accounts for approximately 70% of the total volume. Analyses based on the full corpus must therefore take into account the overrepresentation of Wikipedia.

Each subcorpus (excluding *The Little Prince*) is representative of a genre. This representativeness is achieved not through sampling but through complete collection. The first genre in the corpus is the Wikipedia encyclopedia[5]. It is worth noting that some articles contain simplified and traditional characters, indicating that native speakers from Hong Kong or Taiwan may have contributed. However, the annotators involved in our work did not detect any syntactic mixing; the language used is Mandarin Chinese as spoken in Beijing.

The Wikipedia subcorpus is primarily representative of Wikipedia's own style. We cannot claim that it is representative of encyclopedias in general. However, it is reasonable to assume that Wikipedia shares many features with traditional encyclopedias: a very large vocabulary, an almost complete absence of dialogue and direct speech, and very rare use of personal pronouns and address forms, etc.

The second genre is journalism. This subcorpus consists of news articles scraped from newspaper websites, including national newspapers (for the four languages) and regional newspapers (for Chinese). Using national rather than regional press increases the likelihood of encountering shared themes and vocabulary among the four languages. The collection process involved listing relevant news websites and downloading all available articles at the time of collection.

4) The corpus consists of raw texts but includes information about document structure (headings, paragraphs, etc.). Editorial metadata (such as author names or section names) are excluded from the body text and instead marked with tags, as they do not influence proper name quantification.

5 & 6) CompCor contains only complete texts, and each component (paragraphs, headings, etc.) is annotated.

7) The corpus is partially open-access. Wikipedia and *The Little Prince* are freely available. The annotated versions of these are also open. For the journalistic texts, only the URLs are provided; users must retrieve the articles themselves and convert them to plain text.

The annotated versions are distributed in CoNLL format, indicating the start and end character positions of each token in the article. This setup is similar to that of the BCCWJ (Maekawa et al., 2013): only annotation instructions are provided, not the text itself (see for example `grew.fr/download/SUD_2.16/SUD_Japanese-BCCWJ.tgz`). There is, however, a

[4]Available at `https://sharedocs.huma-num.fr/wl/?id=eeA0Otx6Ma1yorPAK9CI3e4rJy89DArs`

[5]Source: `dumps.wikimedia.org`

difference between the BCCWJ and us: in BCCWJ, the text must be pre-segmented as only token numbers are provided. We prefer to offer data in a format that is independent of segmentation.

Tables 1-2 provide an overview of CompCor. The figures show that the corpus is generally large in size. Korean is smaller due to fewer available resources. The data confirms the overrepresentation of Wikipedia, except in the case of Chinese.

6. Annotation Format

At the current stage or the project, the quantificational annotation must identify QNEs, quantifiers, the relations between them, and coreference chains.

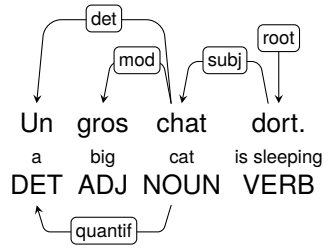
Given the size of the corpus to annotate, the annotation must be automatable. To this end, we rely on syntactic pre-annotation from which at least part of the quantificational annotation can be inferred.

We have chosen a dependency-based syntactic annotation, as it is possible to convert a dependency representation into a constituency-based one, whereas the reverse is only partially true (Kahane, 2012). Among the different types of dependency representations, we have selected the SUD (Surface Syntactic Universal Dependencies) format (Gerdes et al., 2019) because it is explicitly syntax-oriented. In principle, syntactic analyses in SUD are therefore unlikely to interfere (i.e., cause analysis conflicts or redundancies) with our semantic analyses. Another advantage of SUD is that it is well-documented, with many large-scale treebanks available for a wide range of languages[6], including the ones we are working with. This format is currently maintained by a very active community. And finally, SUD is compatible with UD (Universal Dependencies) format, which makes our annotation accessible to a large community of researchers.

In practice, the annotation involves marking the heads of the nominal expressions and the heads of quantifiers. Each quantified QNE's head is then linked to the head of its quantifier. This relation is labeled `quantif`, which should be understood as "is quantified by". Unquantified QNEs are referred to as "bare QNEs". This term is used here in a semantic sense, broader than the syntactic usual expression "bare nouns", which typically refers to a noun without a determiner. Ex.3 illustrates the parsing of a simple sentence. In the graphical representation, links above the annotated expressions correspond to the SUD representation; links below reflect the quantificational annotation; the part-of-speech tags follow SUD conventions.

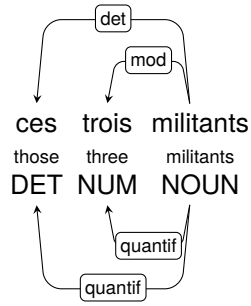
[6]See `surfacesyntacticud.org`

(3)



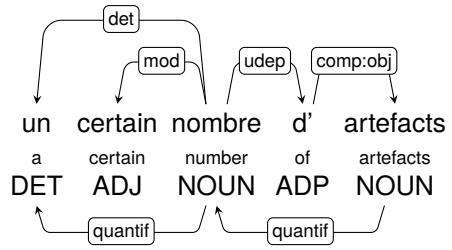
It should be noted that a QNE may have multiple quantifiers, and a quantifier itself may be a noun phrase containing another quantifier.

(4)



The quantifier can itself be a noun phrase containing another quantifier.

(5)



In the CoNLL-format transcription of the SUD representation, quantificational information is provided in two additional columns (see Tab. 3). In the literature —especially concerning CJK languages —token segmentation is a frequently discussed issue. In the raw text of CompCor, the original segmentation is preserved: Chinese and Japanese are unsegmented; Korean and French are segmented according to the natural segmentation present in the source texts.

For quantificational analysis, the raw text is resegmented. French is resegmented following the conventions used in SUD. The only difference is that we do not lemmatize the raw text. Thus, the French determiner *du* is kept in its surface form, rather than being lemmatized as *de le*.

For CJK languages, a two-stage segmentation is performed. In Chinese, the first segmentation follows SUD conventions. For Japanese and Korean, we use two similar segmentations: "Small Unit Word" (SUW) for Japanese (Omura et al., 2021), and MorphUD for Korean (Park and Tyers, 2019). These initial segmentations bring Japanese and Korean analyses closer together.

A second, finer segmentation is then applied. Blin and Choi (2025) demonstrate that SUW and

	Ko	Fr	Ja	Ch	Total
Press article	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
Le Petit Prince	1	1	1	1	4
Total	1 200 145	4 110 002	2 030 946	468 321	7 809 414

Table 1: #Documents per corpus in CompCor corpus

	Ko	Fr	Ja	Ch	Total
Press article	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
Le Petit Prince	115	86	111	76	388
Total	7 818 400	39 888 422	23 003 969	4 463 327	75 174 118

Table 2: #phrases per corpus in CompCor corpus

MorphUD do not allow for fine-grained analysis of Sino-Japanese and Sino-Korean compounds. The same applies to Chinese. In particular, nominal monomorphemes are not systematically tokenized. Therefore, we apply a second segmentation to all semantically transparent tokens composed of a nominal monomorpheme and a quantifying monomorpheme. For example, in Japanese, the token 各市 $_{N}$ *kaku-shi* "each city" is segmented into two tokens: 各 $_{D}$ *kaku* "each" and 市 $_{N}$ *shi* "city". To ensure consistency across analyses, we segment any semantically transparent token that contains at least one nominal monomorpheme. This includes compounds where two nominal morphemes are concatenated, and the first serves as an argument of the second, such as (ja) 水温 *sui-on* "water temperature" or 湯量 *yu-ryō* "hot-water quantity". This segmentation is relevant for both monolingual processing and cross-linguistic comparison. In fact, we observe that in these CJK compounds, the first monomorpheme is never quantified. However, in semantically equivalent genitive constructions in French, the element corresponding to the first CJK morpheme "can" be quantified (comparison with English):

(6) (Fr) cette quantité [d' / *de l'] eau
(En) this quantity [of / *of the] water
(Fr) cette température [*d' / de l'] eau
(En) this temperature [of / *of the] water

7. Experiment

We manually applied the quantificational annotation to a subcorpus of CompCor. This operation serves three purposes. The first is to test our quantificational representation rules on real data. It also allows us to identify potential difficulties (discussed in Section 7.4). The second goal is to use this experience to draft a quantificational annotation guideline. The third objective is to produce a gold-standard corpus, either to evaluate annotation models or to serve as a seed corpus for bootstrapping-based training. Additionally, we use this corpus to compare the frequency of bare QNEs across the four languages, as well as the frequency of different categories of quantifiers.

In what follows, we describe the resource, the method, and the results. We conclude with a discussion of the challenges encountered during the annotation process.

7.1. Corpus

To create the sample, we randomly extracted, from CompCor, eight news articles of varying lengths for each language. In Japanese and Korean, each newspaper is represented by two articles. In Chinese, the sample includes articles from eight different newspapers. Additionally, we include one Wikipedia page and 25 standalone sentences from the GSD Treebank (Google UDT 2.0; McDonald et al., 2013), which is also composed of sentences from Wikipedia. These sentences are available with SUD annotation and are used to assess the compatibility between quantificational annotation and SUD annotation. The sample also includes the title and the first chapter of *The Little Prince*. Statistics for the sample are provided in Table 4. The table does not list token counts for Japanese and Chinese, as the original texts are unsegmented. The Korean and French tokens correspond to "natural" tokens, defined by standard writing conventions. These are not always linguistically motivated (see discussion in Park and Tyers, 2019).

We deliberately chose to include a majority of news articles. This type of text, in our view, is far more frequently read by speakers than encyclopedic texts. This choice allows us, to some extent, to work with language that is closer to what speakers use in their everyday lives.

```
5  ces     ce     DET   _  Number=Plur|PronType=Dem  7  det       _  _  7  quantif
6  trois   trois  NUM   _  ExtPos=DET|Number=Plur    7  mod       _  _  7  quantif
7  points  point  NOUN  _  Gender=Masc|Number=Plur   4  comp:obj  _  _
```

Table 3: CoNLL-formatted file with two additional columns dedicated to quantification-related information.

	Press articles		Wikipedia+GSD		Le Petit Prince	
	sentences	token	sentences	token	sentences	token
Chinese	129		36		37	
French	192	4 445	47	1 066	42	554
Japanese	101		39		49	
Korean	250	3 776	80	887	48	403

Table 4: Sample description; the number of tokens is not provided for Japanese and Chinese, as these languages are not "naturally" segmented.

7.2. Annotation

The annotation process began by identifying and POS-tagging the heads or the QNEs and the heads of the quantifiers. We then added quantificationnal relations in accordance with the annotation guidelines.

We used the Label Studio software (Tkachenko et al., 2020-2025) for annotation, as it does not require pre-segmentation. This gave us full flexibility in choosing the token boundaries. The software also allows for the handling of full texts. For the present task, which is relatively simple, this tool was sufficient. However, it does not support complex annotations well, as the visualization of links becomes unreadable. For instance, it would not be possible to simultaneously display both the SUD syntactic annotation and the quantificational annotation.

An annotation guideline was written. It covers annotation for all four languages and includes around 80 rules, along with approximately the same number of examples. Wherever possible, the rules are designed to be cross-linguistic in scope, avoiding ad hoc treatments. The guide also includes also language-specific rules.

The annotation was carried out manually by senior linguists (faculty members and/or researchers) and PhD students in linguistics or NLP. All annotators are either native speakers of the languages they annotate or have near-native proficiency (minimum C2 level according to the Common European Framework of Reference; Council of Europe, 2001). The annotation team is coordinated by the first author, who designs the guideline and updates it based on annotator feedback. Each text is independently annotated by at least two annotators, who then compare their analyses and report their observations, leading to updates of the guideline. This procedure explains the very high agreement rates, ranging between 80% and 90%.

	Ch	co	Ja	fr
% Bare nouns	91.0	88.4	91.2	23.5
% Types of quantifiers				
Numeral	70.5	40.8	48.3	6.9
Demonstratives	13.7	20.4	12.1	4.1
Distributives	7.1	2.5	1.7	0.1
...				
Indefinites	72.1	43.6	50.0	20.2
Definites	23.0	28.0	22.4	55.4
% Types of quantifiers; including ∅ quantifiers				
Numeral	6.4	4.7	4.3	5.3
Demonstratives	1.2	2.4	1.1	3.1
Distributives	0.6	0.3	0.2	0.1
...				
Indefinites	6.5	5.1	4.4	15.5
Definites	2.1	3.2	2.0	42.4

Table 5: Statistics on QNEs and quantifiers in the sample

7.3. Results

Tab.5 presents the main results of the experiment.

Regarding the proportion of bare QNEs (i.e., without quantifiers), the figures confirm a clear distinction between French on one side and the CJK on the other. In French, quantified QNEs overwhelmingly dominate, whereas bare QNEs are predominant in CJK. As for both the frequency of bare QNEs and the nature of quantifiers, no notable differences are observed among the CJK languages.

The most original finding of this experiment concerns the distribution of definite and indefinite markers. Once again, a clear distinction emerges between CJK and French. When CJK languages use quantificational markers, they rely primarily on indefinites —most often numerals. In contrast, the

majority of quantifiers in French are definites (articles, demonstratives, etc.).

If we assume that bare QNEs are quantified by empty quantifiers, then the frequency of each (non-empty) determiner category is ultimately similar across the four languages. The only persistent result is the distinction between definite and indefinite determiners.

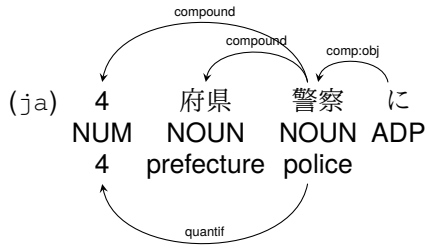
7.4. Discussion

One of the objectives of the experiment was to assess the consistency of the annotation and to identify implementation challenges specific to each language. Here, we return to what we consider to be the most significant issues, particularly those related to the SUD representation.

As previously mentioned, the core principle is to infer all or part of the quantificational representation from the SUD syntactic structure. However, in its current form, this format presents certain limitations and does not always allow for the correct inference of quantificational structures.

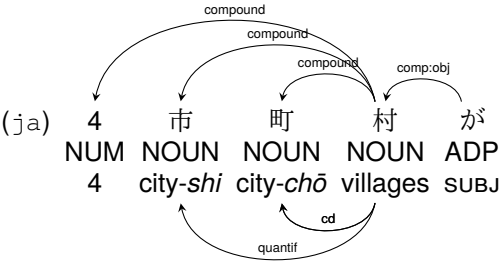
The first issue encountered concerns the representation of coordination and disjunction. Noun phrases formed through concatenation are syntactically represented in the same way, regardless of the semantic relationship between the components. As a result, it is not possible to infer an accurate quantificational representation. For instance, consider Ex.7 and Ex.8. In the first case, the semantic relationship is argumental. In the second case, it is coordination. Consequently, in the first case, the quantifier applies exclusively to the head noun, whereas in the second, it applies to all nouns in the phrase.

(7)

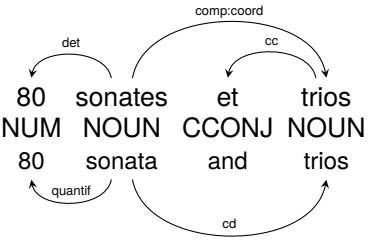


To distinguish between these two modes of quantification, we cannot rely solely on the SUD representation, as it provides a single undifferentiated structure. The quantificational representation must therefore be enriched. We explored several approaches and concluded that the only viable solution is to explicitly introduce, in the quantificational annotation, an additional relation `cd` to indicate coordination or disjunction, as illustrated by the examples (see also the French example 9).

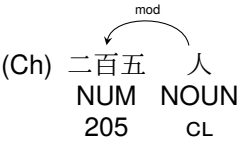
(8)



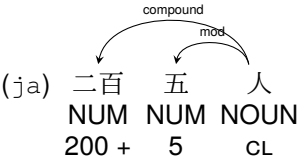
(9)

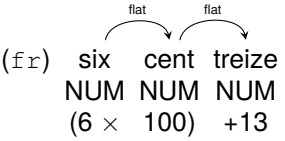


Several other problems arise from clear incompatibilities across SUD representations in different languages. Resolving these issues requires upstream adjustments during the syntactic annotation phase. Such inconsistencies become particularly apparent when they involve structurally identical constructions with the same semantic value. This is notably the case with Sino-origin vocabulary. For example, the prefix (`ch`, `ja`) 各 (`ko`) *gak* "each" is categorized as a determiner in Chinese and Korean, resulting in a `det` dependency to the quantified QNE. However, in Japanese, the same morpheme is categorized as a NOUN, and the dependency to the quantified QNE is marked as `compound`. Ex.10 also shows that Chinese-character numerals (in Chinese and Japanese), and spelled-out numerals in French receive three different treatments. Pending a long-term solution, it seems reasonable to simplify by aligning with the Chinese treatment, where the numeral is handled as a single token.



(10)





8. Conclusions and Future Directions

The present work constitutes the first step in a larger project aimed at building a resource designed to compare modes of quantification of QNEs, in context, across multiple languages and discourse genres. We have outlined the corpus selection criteria and presented the quantificational annotation project.

Given the challenges encountered with syntactic annotation (see previous section), it was not possible to formulate inference rules to derive the quantificational representation from the syntactic representation. To achieve this, it has become clear that the next phase of the project must focus on unifying SUD syntactic descriptions —at least for constructions involved in quantification.

Moreover, the experiment showed that even for seemingly identical structures —such as Sino-origin vocabulary —annotators struggled to arrive at compatible descriptions. This significantly slows down the annotation process. Therefore, in the next phase, it would be advisable to narrow the scope of observation and focus on a specific subtype of quantifiers. We are particularly considering demonstratives, as they are limited in number.

Despite the difficulties we have discussed in this text, new studies show encouraging results and confirm the validity of quantificational annotation. We have applied this annotation to French-language corpora of French-Korean bilingual children. The language involved is "simple": short sentences, limited vocabulary, and quantificational expressions reduced to numerals or determiners. Under these conditions, quantificational annotation carried out by an automatic annotation based on deterministic rules proceeds successfully. It is therefore already possible to apply the annotation to large volumes of linguistically simple texts.

9. Bibliographical References

Steven P Abney. 1987. *The English noun phrase in its sentential aspect*. Ph.D. thesis, Massachusetts Institute of Technology.

J. Barwise and R. Cooper. 1981. Generalized quantifiers and natural language. In J. Kulas, J.H. Fetzer, and T.L. Rankin, editors, *Philosophy, Language, and Artificial Intelligence. Studies in Cognitive Systems*. Springer.

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

Raoul Blin and Jinnam Choi. 2025. Segmentation of sino-origin words to enhance the representation of korean and japanese in s/ud-format treebanks. In *Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, Ljubjana,Lituania.

Raoul Blin, Alexander Delaporte, Ilaine Wang, Arnaud Arslangul, Xinyue Cécilia Yu, and Camille Noûs. 2025. Compcor-v0. 1.3: Corpus comparable coréen français japonais mandarin.

Francis Bond. 2001. *Determiners and Number in English contrasted with Japanese, as exemplified in Machine Translation*. Ph.D. thesis, University of Queensland.

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

Jacee Cho and Roumyana Slabakova. 2014. Interpreting definiteness in a second language without articles: The case of l2 russian. *Second language research*, 30(2):159–190.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2019. Improving surface-syntactic Universal Dependencies (SUD): MWEs and deep syntactic features. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 126–132, Paris, France. Association for Computational Linguistics.

Jena D. Hwang, Hanwool Choe, Na-Rae Han, and Nathan Schneider. 2020. K-SNACS: Annotating Korean adposition semantics. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 53–66, Barcelona Spain (online). Association for Computational Linguistics.

Sylvain Kahane. 2012. Why to choose dependency rather than constituency for syntax: a formal point of view. In *Meaning, Text, and other Exciting Things. A Festschrift to Commemorate the 80th Anniversary of Professor Igor Alexandrovic Mel'cuk*, page 257-272. Languages of Slavic Culture.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2013. Balanced corpus of contemporary written japanese. *Language Resources and Evaluation*, 48(2):345–371.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Mai Omura, Aya Wakasa, and Masayuki Asahara. 2021. Word delimitation issues in ud japanese. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 142–150.

Jungyeul Park and Francis Tyers. 2019. A new annotation scheme for the Sejong part-of-speech tagged corpus. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 195–202, Florence, Italy. Association for Computational Linguistics.

Siyao Peng, Yang Liu, Yilun Zhu, Austin Blodgett, Yushi Zhao, and Nathan Schneider. 2020. A corpus of adpositional supersenses for Mandarin Chinese. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5986–5994, Marseille, France. European Language Resources Association.

Lisa Raithel, Hui-Syuan Yeh, Shuntaro Yada, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Philippe Thomas, Tomohiro Nishiyama, Sebastian Möller, et al. 2024. A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions across Languages;. pages 395–414.

Nils Reimers and Iryna Gurevych. 2020. 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.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel & free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.