J-SNACS: Adposition and Case Supersenses for Japanese *Joshi*

Tatsuya Aoyama[♢], **Chihiro Taguchi**[♠], **Nathan Schneider**[♢]
[♢]Georgetown University, [♠]University of Notre Dame
{ta571, nathan.schneider}@georgetown.edu, ctaguchi@nd.edu

Abstract

Many languages use adpositions (prepositions or postpositions) to mark a variety of semantic relations, with different languages exhibiting both commonalities and idiosyncrasies in the relations grouped under the same lexeme. We present the first Japanese extension of the SNACS framework (Schneider et al., 2018), which has served as the basis for annotating adpositions in corpora from several languages. After establishing which of the set of particles (*joshi*) in Japanese qualify as case markers and adpositions as defined in SNACS, we annotate 10 chapters (≈10k tokens) of the Japanese translation of *Le Petit Prince* (The Little Prince), achieving high inter-annotator agreement. We find that, while a majority of the particles and their uses are captured by the existing and extended SNACS annotation guidelines from the previous work, some unique cases were observed. We also conduct experiments investigating the cross-lingual similarity of adposition and case marker supersenses, showing that the language-agnostic SNACS framework captures similarities not clearly observed in multilingual embedding space.

Keywords: semantics, adpositions, function words, similarity metrics, Japanese

1. Introduction

Adpositions and case markers—such as the English prepositions *in*, *for*, and *through*—play a significant role in constructing the overall meaning of a given sentence (e.g., *I bought the bike **for** him* vs. *I bought the bike **from** him*), yet they are highly polysemous and contextual (e.g., *I bought the bike **for** him* vs. *I rode the bike **for** an hour*).[1] To capture these properties of adpositions and case markers, Schneider et al. (2018) proposed a comprehensive and largely language-agnostic annotation scheme called the Semantic Network of Adposition and Case Supersenses (SNACS). Several corpora have been annotated with this framework (e.g., Kranzlein et al., 2020), and it has been applied and extended to a number of typologically different languages, such as German (Prange and Schneider, 2021), Korean (Hwang et al., 2020), and Hindi (Arora et al., 2022), among others.

Here we present "J-SNACS", the first extension of the SNACS framework to Japanese, which features a rich system of *joshi* (particles; §2). Our main contributions are:

- criteria defining which particles should be considered annotation targets (§3);
- annotations of particles in 10 chapters (≈10k words) of a Japanese translation of *Le Petit Prince*, amounting to 1,810 annotated targets (§5);
- an annotation study demonstrating high agreement (§4); and
- an analysis of the cross-lingual similarities of adpositions and case markers using supersense distributions and contextualized word embeddings (§6).

Overall, we find that the extended SNACS framework was necessary to account for some of the Japanese particles, such as topic and focus markers, as was the case with Korean (Hwang et al., 2020). We also find that the construal analysis provides a powerful device to capture the subtle differences in the way certain particles carry nuanced meanings, as exemplified by the quotative particle *to* (§5.3.3). Lastly, our experiments show that, while similarities of adpositions and case markers in Japanese and English are encoded as multilingual embedding distance to some extent, the SNACS supersense distributions facilitate more meaningful cross-lingual comparisons. We make our data and code available online.[2]

2. Related Work

2.1. SNACS

The SNACS framework categorizes the meaning of adpositions and case markers into 52 semantic classes called **supersenses**, such as Time (e.g., *class **on** Friday*) and Locus (e.g., *restaurant **in** DC*). Except for a few language-specific extensions, these supersense labels are shared across different languages, including ones that are never realized as adpositions in English (e.g., Content). On top of this language-agnostic repository of supersenses, the SNACS framework adopts a **construal** analysis (Hwang et al., 2017), which distinguishes the meaning a given adposition or case marker conveys in a given context from the meaning it contributes on its own. Both the contextual and

[1] See §2.2 for more details on terminology.

[2] https://github.com/t-aoyam/japanese-snacs

LREC-COLING 2024, pages 9604–9614
20-25 May, 2024.

non-contextual meanings are drawn from the same supersense repository, and the two parts of the construal analysis are called **scene role** and **function**, respectively, denoted as SCENEROLE⇝FUNCTION. For example, an English adposition *for* often has BENEFICIARY as its prototypical function, yet the scene role may vary from context to context:

(1) It's a gift **for**/BENEFICIARY Tom.

(2) It's sad **for**/EXPERIENCER⇝BENEFICIARY Tom.

In (1), the adposition *for*, in conjunction with the overall context, construes Tom as the beneficiary of the gift, which is congruent with the meaning contributed by *for* alone. In (2), whereas the non-contextual meaning of *for* is BENEFICIARY, in context, Tom is an EXPERIENCER of the sadness, hence the construal EXPERIENCER⇝BENEFICIARY.

With this powerful construal analysis and the shared inventory of 52 supersenses,[3] SNACS has been extended to typologically diverse languages, such as Chinese (Peng et al., 2020), Finnish and Latin (Chen and Hulden, 2022), German (Prange and Schneider, 2021), Gujarati (Mehta and Srikumar, 2023), Hindi (Arora et al., 2022), and Korean (Hwang et al., 2020). However, it has not yet been applied to Japanese. This study is the first such extension.

2.2. Japanese *Joshi* (particles)

Japanese is a postpositional language, and the treatment of postpositions and case markers in Japanese linguistics varies widely. These two categories are often lumped together under the umbrella term *joshi*, which is roughly translated as *particles*. For example, Kawashima (1999) defines Japanese particles with criteria that they (1) show their relationships to other words and/or give other words a particular meaning, (2) do not inflect, (3) correspond to prepositions, conjunctions, and interjections in English, and (4) are placed after the word they modify. Kawashima (1999) also lists 119 items under the category *particle*, with no further subcategorization. Similarly, Chino (1991) defines the usages of 69 particles, calling 16 of them *sentence final particles*, and the rest simply *particles*. Some Japanese grammar books also put both case markers and postpositions under *particles* (e.g., Akiyama and Akiyama, 2012; Sato, 2021).

Others, however, adopt a finer-grained categorization. For example, McGloin et al. (2014) distinguish *case particles*, *postpositions*, *adverbial particles*, and *sentence-final particles*. Similarly, Siegel's (1999) taxonomy includes 5 types. Among the finest-grained is that of Kaiser et al. (2003), where particles are grouped into 6 top-level categories, with one of them further branching out to 7 subtypes.

Den et al. (2007) developed `UniDic`, a unified taxonomy of Japanese parts of speech with the aim of facilitating morphological analyses for NLP. They categorized Japanese particles into six types: `particle (case)`, `particle (binding)`, `particle (conjunctive)`, `particle (nominal)`, `particle (phrase-final)`, and `particle (adverbial)`.[4] In Japanese Universal Dependencies project (JUD), Tanaka et al. (2016) maps these six categories onto the following four Universal POS (UPOS) tags: `ADP`, `CONJ`,[5] `SCONJ`, and `PART`, which is in line with the above-mentioned definition of Japanese particles in Kawashima (1999).

3. Scope of Investigation

3.1. Insights from XPOS and UPOS

Since the aforementioned XPOS-UPOS mapping is many-to-many, and XPOS and UPOS seem to capture slightly different aspects of various particles' syntactic status, both can help us in disambiguating what qualifies as an annotation target for SNACS. As such, we define our annotation targets based on the combination of XPOS, UPOS, and lemma, guided by the principles of SNACS (Schneider et al., 2022), in which an adposition:

(3) a. mediates a semantically asymmetric figure-ground relation between two concepts, and
b. is a grammatical item that can mark an NP.

Note that (3b) encompasses certain lexical items that mark NPs even where they are used to mark clauses (as a subordinator) or are intransitive.[6] We also refer to SNACS guidelines for other languages, such as Korean (Hwang et al., 2020) and Mandarin Chinese (Peng et al., 2020).

First, we provide descriptive statistics from a Japanese corpus annotated for both XPOS and UPOS. We use Japanese-GSD v2.6,[7] which contains sentences originally from Google Universal Dependency Treebanks v2.0 (McDonald et al., 2013). This Japanese corpus contains 8,100 sentences from news and blog articles,[8] totaling ≈200k tokens. In this corpus, XPOS tags are based on

[3] See `http://www.xposition.org/supersenses`.

[4] We use a `monospace` font for UPOS and XPOS.

[5] Renamed to `CCONJ` in UD v2 (Nivre et al., 2020).

[6] E.g., **through** is considered an adposition per (3) in *We broke [**through** the wall]* (an NP-marking preposition) as well as *We broke **through*** and ***Through** faking our identities, we managed to escape*.

[7] Data available at `https://github.com/UniversalDependencies/UD_Japanese-GSD`.

[8] See `https://universaldependencies.org/treebanks/ja_gsd/index.html` for the details.

XPOS	UPOS	lemmas	
particle (case)	ADP	の(8882), に(6429), を(5340), が(4117), と(3846)	✓
	SCONJ	に(104), の(38), で(13)	✓
	CCONJ	で(23), に(2)	✗
particle (binding)	ADP	は(5542), も(1844), こそ(16)	✓
	SCONJ	も(20), は(5)	✓
particle (nominal)	SCONJ	の(842)	✗
particle (conjunctive)	SCONJ	て(5258), が(784), と(270), ば(143), ながら(76)	✗
particle (adverbial)	ADP	や(610), など(453), まで(286), か(182), だけ(100)	✓
	PART	か(96), など(78), たり(76), だけ(35), ほど(27)	?
particle (phrase-final)	PART	か(146), よ(57), ね(57), な(36), わ(4)	✗

Table 1: XPOS to UPOS mapping of Japanese particles. ✓ represents a combination of XPOS and UPOS that is unambiguously included as annotation targets; ✗ represents unambiguous exclusion; and **?** represents a combination of XPOS and UPOS whose inclusion is lemma-dependent.

UniDic (Den et al., 2007), which were then manually converted into UPOS tags as defined in JUD (Tanaka et al., 2016; Asahara et al., 2018).[9]

Table 1 illustrates the mapping of Japanese UniDic-based XPOS tags to UPOS tags, as well as the top 5 respective lemmas appearing in the corpus. First, particle (phrase-final), particle (nominal), and particle (conjunctive) only map to PART, SCONJ, and SCONJ, respectively. Following Peng et al. (2020), who excluded phrase-final particles from SNACS annotation targets for Mandarin Chinese, we will also exclude them. For particle (nominal), the particle の*(no)* and its phonological variant ん*(N)* are both used as a complementizer (see (4a) in §3.2), and are hence excluded as annotation targets. For particle (conjunctive), although the rule (3b) stipulates that the particles that *can* mark an NP should be included where it is marking a clause, all instances of particle (conjunctive) were excluded because they behave distinctively differently from when they mark an NP. For example, the usages of で*(de)* as particle (case) and particle (conjunctive) are compared in (5a) and (5b) (see §3.2).

A similar criterion based on rule (3a) was applied to particle (adverbial). For example, as まで*(made)*, which translates to *by* or *until*, is commonly used to mark an NP as ADP, its usage as PART is also included. Among the excluded ones from particle (adverbial) is たり*(tari)*, as it is never used to mark an NP. For particle (case), only the ones used as ADP or SCONJ were included and the ones annotated as CCONJ were excluded, as they only occur sentence-initially without a clear reference to two concepts, violating rule (3a).

Lastly, particle (binding) is called *binding*, based on its unique behavior in classical Japanese, where the subsequent verb conjugation is decided (i.e. bound) by the presence of this particle. In modern Japanese, more frequent particles of this kind include topical and focus particles, such as は(ha) and も(mo). Following Korean SNACS (Hwang et al., 2020), which included topical and focus particles, we also include them as annotation targets.

[9]Detailed guidelines available at http://fginter.github.io/docs/ja/pos/all.html.

3.2. Ambiguity in Identifying Annotation Targets

Introduced below are examples of annotation targets and non-targets for particles of the same surface form. Consider the ambiguity of the particle の*(no)*, illustrated in (4):

(4) a. koosu-ga kanari yasui-**no**-ga ...
course-NOM very cheap-**PRT**-NOM ...
That the course is very cheap is ... (*train-s4781*)

b. kochira-ga saisho-**no** mise .
this-NOM first-**GEN** store .
This is the first store. (*train-s4857*)

In (4a), the token の*(no)* functions as a particle (nominal), which can be roughly translated as an English complementizer *that*. In (4b), the same surface form の*(no)* functions as a particle (case), marking the preceding noun as genitive. Therefore, the former is not an annotation target, whereas the latter is, mirroring the inclusion of possessive markers in English SNACS (Blodgett and Schneider, 2018; Schneider et al., 2022).

(5) a. bijinesu-**de$_1$** kakawaru ue-**de$_2$** ,
business-**PRT** interact on-**PRT** ,
When cooperating **in** business, (*train-s6432*)

b. nan-do yon-**de**-mo ko-nai .
what-time call-**PRT**-FOC come-NEG .
(They) don't come no matter how many times (I) call (them). (*train-s2734*)

In (5a), で$_1$ (de$_1$) is used as ADP, and is clearly marking the preceding noun, whereas で$_2$ (de$_2$), which

Phase	# P	Target			Raw Agreement			Kappa		
		P	R	F	SR	Fxn	SR⇝Fxn	SR	Fxn	SR⇝Fxn
1	443	.97	.94	.95	.51	.64	.40	.54	.67	.42
2	483	.98 +.01	.92 −.02	.95	.68 +.17	.77 +.13	.63 +.23	.73 +.19	.84 +.17	.69 +.27

Table 2: Inter-annotator agreement scores at two phases of annotation: number of particles (#P); precision, recall, and F1 of annotation targets; and raw agreement rate and Cohen's kappa, each reported just for scene role supersenses (SR), just for function supersenses (Fxn), and for their combination.

is annotated as SCONJ, is idiomatically forming a subordinate-conjunctive phrase with the preceding word *ue*. Although de$_2$ is semantically bleached, its original function in this context is still the marking of the preceding noun as locative. Compare this to (5b), where *で(de)* is simply adding a conjunctive meaning to the preceding verb, without marking an NP at all. Hence, usages of *で(de)* in (5a) are included, while the one in (5b) is excluded.

4. Corpus Annotation

4.1. Data and Preprocessing

Although multiple Japanese translations of *Le Petit Prince* exist, to the best of our knowledge, only the version we used in our study[10] has been made publicly available with the CC-BY license (Okubo, 2014). This version is a direct translation from the original language (French) to Japanese. We used spaCy[11] with ja-ginza-electra[12] for sentence segmentation, tokenization, and POS tagging.

The data were then converted into .xlsx format for the subsequent annotation rounds, with each row containing each token in Hiragana, Kanji, and Roman characters, as well as its UPOS and XPOS tags. The first two authors, both of them Japanese native speakers and Ph.D. students in computational linguistics, served as annotators. All annotated data were converted into the standardized .conllulex[13] format prior to any data analyses or experiments, described later in this paper.

4.2. Inter-Annotator Agreement

We first conducted a practice annotation session on English examples to familiarize ourselves with the general SNACS guidelines. This phase involved annotating the English Little Prince corpus and checking if our annotation decisions were in line with the gold labels.

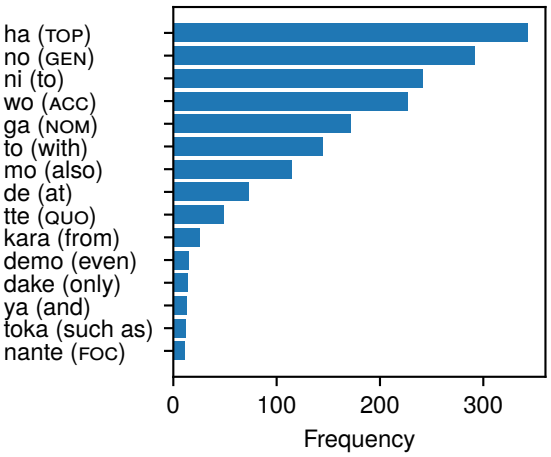

Figure 1: Frequency breakdown by word type of the 15 most common particles.

We then moved on to the Japanese Little Prince corpus, and started the identification of the annotation targets, as well as the supersense annotation. To measure and improve the quality of annotation, we first annotated the first 3 chapters independently (phase 1), and subsequently conducted a thorough adjudication to minimize the disagreements. We then annotated the next 3 chapters (chapters 4–6) independently (phase 2). Chapters 7–10 were annotated by Annotator 1 alone.

Table 2 summarizes the agreement between the 2 annotators. Across phase 1 and phase 2, the identification of annotation targets (i.e. particles) is consistent between the two annotators, with the F1 score staying sufficiently high at .95. Precision, recall, and F1 scores were calculated using Annotator 1 as the "gold" standard. For the raw agreement and Cohen's Kappa, the improvement between the two phases is substantial across all metrics. The post-adjudication (phase 2) Kappa scores were considered sufficiently high and representative of the data (≈25% of the data), although there still seems to be room for improvement.

5. J-SNACS in Action

5.1. Descriptive Statistics

Figure 1 shows counts of the 15 most frequent particles. The most frequent particle was the topic

[10] https://www.aozora.gr.jp/cards/001265/files/46817_24670.html

[11] https://spacy.io/

[12] https://github.com/megagonlabs/ginza

[13] converter available in our repo: https://github.com/t-aoyam/japanese-snacs/blob/main/code/xlsx2conllulex.py

Count		Type-Level Frequency	
Chapters	10	Scene Role	49
Sentences	619	Function	40
Tokens	9,951	SR⇝Fxn	135
Annotation Targets	1,810	SR = Fxn	38
		Particles	30

Table 3: Descriptive statistics of the corpus. Left columns represent count data, and right columns represent type-level frequencies.

marker は(ha), which accounts for about 18% of all the particles in the corpus. In fact, case markers, namely は(ha), を(wo), and が(ga), add up to almost 40% of the tokens. As Korean has similar particles, we refer to the Korean adaptation of SNACS (Hwang et al., 2020).

Table 3 summarizes the basic statistics of the data, as well as of the distribution of supersenses. Notably, of the 9,951 word tokens in the first 10 chapters of *The Little Prince*, 1,810 of them are particles, accounting for about 18% of the tokens. In contrast, the number of unique annotation targets (i.e., particle types) is 30, which is much fewer than the 60 unique adpositions in the English Little Prince corpus (Schneider et al., 2018), but around the same as the Korean counterpart of 29 (Hwang et al., 2020). (6) gives insight into why Japanese has a large number of particle *counts* with a small number of particle *types*:

(6) hon-ni/LOCUS-ha/TOPICAL kaka-re-teita
book-DAT-TOP write-PASS-ING
in the book, it was written... (*ch0-s3*)

Here, whereas the English translation (*in the book*) only requires one adposition, the entire adpositional phrase is topicalized in Japanese, adding an extra annotation target (*as for what was written in the book*). Stacking multiple adpositions and case markers is by no means uncommon in Japanese, and the myriad of such cases in our data explains the lower number of particle *types* and the higher number of particle *tokens*.

5.2. Polysemy in Japanese Particles

Figure 2 summarizes the number of distinct construal pairs attested for each of the 15 most polysemous particles. The particles に(*ni*) and の(*no*) are shown to be the most polysemous in our dataset, with both having 37 unique construal pairs. The high degree of polysemy observed for the dative particle に(*ni*) is unsurprising, given that it has many English translations, such as *to*, *for*, *in*, and *at*. Four of the many usages of this particle appear in (7):

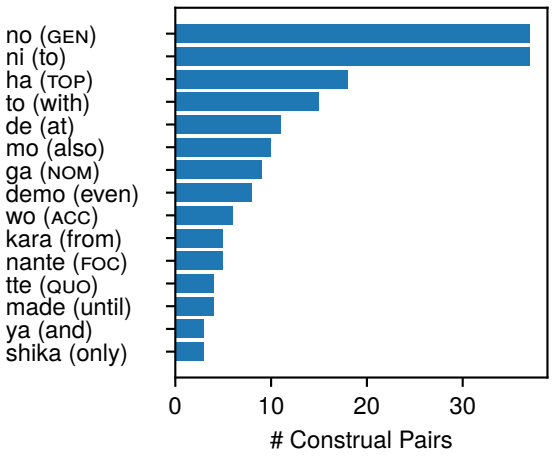

Figure 2: Number of distinct construal pairs for the 15 most polysemous particles.

(7) a. boku-**ni**/BENEFICIARY⇝GOAL hitsuji-no
I-**DAT** sheep-GEN
e-wo kai-te.
picture-ACC draw-IMP.
draw me a sheep (*ch2-s13*)

b. mata aru hi-**ni**/TIME-ha
again one day-**DAT**-TOP
on another day (*ch5-s40*)

c. tora-nante, boku-no hoshi-**ni**/LOCUS-ha
tiger-FOC I-GEN planet-**DAT**-TOP
i-nai-yo
exist-NEG-PRT
of course there is no such thing as tiger on my planet (*ch8-s33*)

d. hi-**ni**/SETITERATION hi-ni dandan
day-**DAT** day-DAT gradually
wakat-te ki-ta
understand-PRT come-PAST
came to gradually understand day by day (*ch5-s0*)

In (7a), *ni* is indicating the potential RECIPIENT or BENEFICIARY of the drawing. This is an extended usage of the typical use of *ni* as a dative marker as in *X-**ni** iku* (go **to** X), hence annotated as GOAL for its function. In (7b, 7c), *ni* is marking the time and place, respectively. In (7d), *ni* is part of an idiomatic expression *hinihini* (day-by-day) describing a gradual and consistent change in state, and is construed as SETITERATION.

These disparate uses of a single particle are interesting, yet each of these uses somehow corresponds to an English preposition. But this is not true of all particles. Next, we turn to ones lacking a straightforward prepositional counterpart in English.

5.3. Noteworthy Particles

In our annotation, we encountered a number of particles exhibiting interesting properties.

5.3.1. Various Scene Roles for Topical *Ha*

The most frequent particle, は(ha), seems to play a variety of scene roles. This is in contrast with the annotation decision of a topic postposition (*-eun*) adopted in Hwang et al. (2020), where it was unambiguously annotated as TOPICAL for both scene role and function. Consider the following usages of the Japanese particle は(ha):

(8) a. boku-**ha**/AGENT⇝TOPICAL dogu-wo
I-**TOP** tool-ACC
tebanashi-ta
release-PAST
As for me, (I) dropped the tool (*ch2-s27*)

b. boku-**ha**/ORIGINATOR⇝TOPICAL
I-**TOP**
sono-ko-ni koe-wo kake-ta
that-boy-DAT voice-ACC give-PAST
as for me, (I) talked to that boy (*ch1-s11*)

c. tashizan-no hoka-**ha**/TOPICAL
addition-PRT outside-**TOP**
besides (arithmetic) addition (*ch7-s44*)

d. anmari tooku-he-**ha**/FOCUS ike-nai
excessively far-DAT-**TOP** go-NEG
cannot go too far (*ch3-s47*)

In (8a), the topic particle は(ha) functions as TOPICAL by re-introducing the familiar entity *this person*; at the same time, however, the re-introduced entity is also the subject of the main predicate *to release*. One may argue that, as in the English translation, there is an invisible pronoun *I* implicitly inserted as the subject of the main predicate; however, although this insertion leads to a well-formed sentence in English, this is not the case in Japanese. In this sense, we argue that the topic particle retains its original role as a topic marker for its function, but for the specific meaning in context (i.e., scene role), it conveys the meaning of AGENT. In a similar vein, in (8b), the topic particle は(ha) introduces the meaning of ORIGINATOR, as the ha-marked entity (*boku*) is the one that is talking to "that boy".

On the other hand, (8c, 8d) show different usages. In (8c), including the continuation not shown above, it can be translated as 'as for (something else) other than arithmetic addition, adults cannot do anything,' and the topic particle は(ha) has no place in the main clause. In other words, it only introduces the topic of *something else other than arithmetic addition*, and hence is annotated simply as TOPICAL for both scene role and function. In (8d), the use of the topic particle is completely optional, and the *ha*-marked adverb *tooku-he* (to a faraway place) is unlikely to be introduced as a topic. Here, *ha* functions as a focus particle marking the contrast. Specifically, in (8d), *ha* is emphasizing the fact that it is a faraway place that one cannot go, as opposed to a nearby place. Therefore, this type of contrastive use of the topic particle は(*ha*) is annotated as FOCUS for both scene role and function.

5.3.2. Genitive *No*

In §5.2, we discussed the highly polysemous particle に(*ni*); in fact, the genitive particle の(*no*) is the other most polysemous particle in this corpus, attested with as many as 37 construal pairs.[14] Having the second highest token count after the topical particle は(*ha*), this genitive particle displays a surprising range of usage, with the prototypical functions including GESTALT and CHARACTERISTIC. For brevity, we will only introduce non-prototypical usages involving THEME and AGENT:

(9) a. hikouki-**no**/THEME⇝GESTALT soujuu
airplane-**GEN** control.
flying of an airplane (*ch1-s24*)

b. ano-ko-**no**/AGENT sumu hoshi-ha
that-boy-**GEN** live planet-TOP
The planet that that boy lives on (*ch2-s2*)

In (9a), scene role and function are clearly different: given the context, *airplane* is an object of *control*, indicating THEME as scene role; on the more literal level, *airplane's control* seems to suggest that the word *control* is part of all the attributes, actions, or characteristics associated with the word *airplane*, justifying GESTALT as function.

On the other hand, in (9b), の(*no*) is within a relative clause (*that that boy lives on*), where the genitive particle almost entirely loses its possession-related meaning, and this use of の(*no*) is identical to the less marked nominative particle が(*ga*). This suggests that, given sufficient data, the range of construals covered by の(*no*) becomes a proper superset of that of が(*ga*), explaining its status as the most polysemous particle in this corpus.

5.3.3. Quotative *To*

As the last example of this non-exhaustive list of unique Japanese particles, we introduce one of the quotative particles と(*to*). This roughly corresponds to an English complementizer *that*, which is not an annotation target due to its syntactic status. In fact, an adposition-like quotative particle is present in Korean as well (*-go*), and to cover the supersense of such particles, Hwang et al. (2020) added a new

[14] Some argue that the genitive particle *no* is semantically undefined, rather than polysemous. See Okutsu (1978) for a comprehensive discussion on this topic.

supersense at the function level, namely QUOTE. Interestingly, however, Japanese quotative particles do not necessarily accompany report verbs:

(10) a. mottomorashii-**to**/CONTENT⇝QUOTE
likely-**QUO**
omou
think
think that it's likely (*ch4-s34*)

b. tekuteku-**to**/MANNER⇝QUOTE isu-wo
trektrek(onomatopoeia)-**QUO** chair-ACC
motte aruke-ba
hold walk-if
if you hold the chair and walk step after step (*ch6-s17*)

c. yukkuri-**to**/MANNER ayashi-ta
slow-**QUO** placate-PAST
placated calmly (*ch7-s70*)

The use of と(*to*) as in (10a) is highly frequent, often appearing in combination with reporting or cognition verbs (e.g., *say*, *think*). In Korean SNACS, this usage of quotative particle is construed as TOPIC⇝QUOTE, where TOPIC marks the topic of what is being communicated through the verb. However, a supersense CONTENT has been newly added in the latest SNACS guidelines (Schneider et al., 2022) as a way to distinguish the actual content of the information being conveyed (i.e., CONTENT) from its main topic (i.e., TOPIC); with this addition, this use of quotative particles in Korean and Japanese as shown in (10a), is best construed as CONTENT⇝QUOTE.

In (10b), the quotative particle is still "quoting" the sound a certain action makes (i.e., onomatopoeia); however, the sound cannot be considered the content or topic of the verb, as *aruk-* (*to walk*) is not a communication or cognition verb. We use MANNER⇝QUOTE construal for this use of *to*. Lastly, in (10c), the *to*-marked element is not a quoted content in any way. Since the QUOTE function is no longer retained in this use, we simply annotate it as MANNER for both scene role and function.

6. Similarities of Particles within and across Languages

Next, we conduct experiments to investigate whether the language-agnostic nature of SNACS provides utility in measuring semantic similarities of adpositions and case markers within and across languages. As a proof of concept, we use SNACS-annotated English and Japanese translations of *Le Petit Prince* for the remainder of this section. The English corpus is freely available online.[15] Since the two corpora are from translations of the same book, we believe that this setup minimizes differences due to domain, topic, and genre, making for a focused cross-linguistic comparison of adposition/-particle behavior. This comparison is performed at a macro level, as our two samples are not entirely parallel (they cover different subsets of chapters of the text) and we do not have sentence or token alignments even for the portions that overlap.

6.1. Experimental Setup

We first measure the supersense-based (SS-based) similarity for all possible pairs of particles. As a sanity check, we will also measure the distance between all possible pairs of particles in embedding space using contextualized word embeddings (CWEs) obtained from a language model, and compute Pearson correlations between the SS-based measure and the cosine similarity in embedding space (CWE-based measure). We hypothesize that a supersense distribution provides a more robust measure of semantic similarity because it abstracts away from the noise present in the context surrounding a particle, while preserving the necessary contextual meaning through the supersense. Hence, we will expect a moderate correlation between the two measures.

More concretely, for the SS-based measure, we follow the steps below:

1. for each particle *type*, obtain a vector of its relative supersense frequencies (a vector of length S, where S is the number of all supersenses);
2. for all possible pairs of particles, compute the Jensen-Shannon divergence.

For 2, we choose Jensen-Shannon divergence because of its symmetric and bounded properties, as opposed to other divergence measures such as Kullback-Leibler divergence.

For the CWE-based measure, we follow the steps below:

1. for each occurrence of each particle, feed to a language model the entire sentence in which it occurs and obtain the particle's CWE;
2. for each particle *type*, average across all the CWEs and obtain a type-level mean CWE;
3. for all possible pairs of particles, compute the cosine similarity between their CWEs.

In 1, we use `bert-base-multilingual-uncased` for all experiments.[16] For multi-word expressions (MWE) or single words that correspond to multiple subwords, we use avereage pooling to obtain a single 768-dimensional CWE. There are various ways of obtaining cross-lingual embedding distances,

[15] https://github.com/nert-nlp/English-Little-Prince-SNACS (version as of Nov. 23, 2023, with SNACS v2.5 annotations of chapters 2, 3, and 6–17)

[16] We also experimented with monolingual models; however, the results were not substantially different.

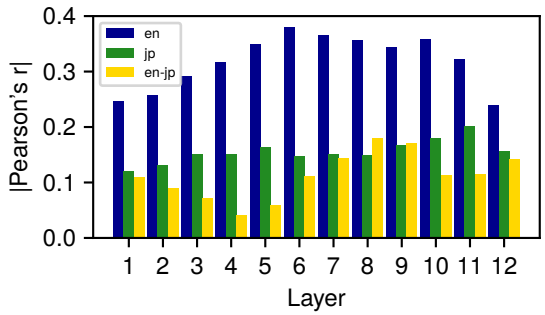

Figure 3: Absolute values of Pearson correlation coefficient *r*'s between SS-based and CWE-based measures, both within and across English and Japanese. Note that all coefficients are negative as they measure the correlations between divergence (SS-based) and similarity (CWE-based).

such as using two separate static or contextual embeddings and aligning them at the word level (Artetxe et al., 2017, 2018). However, such cross-lingual alignment is beyond the scope of this study, and we use mBERT, which Cao et al. (2020) find "somewhat aligned out-of-the-box" (p. 1).

6.2. Results

6.2.1. Within Languages

Figure 3 summarizes the correlations between SS-based and CWE-based measures. For English, with 80 unique particles, the number of comparisons amounted to ${}_{80}C_2 = 3160$ for each of the 2 measures. The correlations between the 2 measures were moderate (−0.25 to −0.4) as expected, statistically significant at all layers ($p < 1e{-}40$).

For Japanese, with the much lower particle type frequency of 30, ${}_{30}C_2 = 435$ pairwise comparisons were conducted for each of the 2 measures. Interestingly, the magnitude of correlations was much lower than English (−0.1 to −0.2) while statistically significant at all layers ($p < 0.01$). One possible reason is that English has more than twice as many particle types (80) as Japanese (30), which makes the mean CWE more discriminating (less heterogeneous). In other words, on average, the identity of the particle will convey more information in English, which may explain the stronger correlations of semantic distances.

6.2.2. Across Languages

The SS-based measure is deemed particularly valuable for cross-lingual settings, because the cross-lingual word alignment in multilingual embedding space has been shown to be weaker for typologically distant languages (Pires et al., 2019; Cao et al., 2020). Hence, we expect the Pearson correlation to be even lower in this setting. Following the same steps for CWE-based and SS-based measures described in §6.2.1, the number of pairwise comparisons amounted to $80 \times 30 = 2400$. As shown in Figure 3, the overall correlation strength is somewhat lower compared to within-language settings, failing to reach statistical significance at layer 4 at $\alpha = 0.01$ ($p = 0.05$). This suggests that the SS-based metric is capturing something different from what the CWE-based metric is capturing.

So far, we have only established that the two metrics are different from each other; here, by manually inspecting pairs that are deemed similar by each of the two metrics, we show that the SS-based metric is indeed more nuanced in capturing the semantic similarities of adpositions and case markers.

Table 4 summarizes the 15 most similar cross-lingual pairs of adpositions and case markers obtained separately from each of the 2 metrics. To add objectivity and systematicity to the evaluation of these pairs, we adopt the following 2 criteria: (1) correspondence with a dictionary translation[17] (**boldfaced**) and (2) conceptual congruence with different polarity or specificity (underlined). For (1), it was considered as satisfying the criterion if, for a given English adposition (first element of each pair), the proposed particle (second element of each pair) is either (i) identical to, or (ii) the rightmost element of, the Japanese dictionary translation. For example, for English adposition *into*, the proposed Japanese particle is *he*, whereas the dictionary translation is *no-naka-he* (of-inside-to; "to the inside of"); however, based on criterion (ii), this was considered to satisfy criterion (1). This is reasonable, given that the rightmost element often decides the overall meaning in Japanese, and that no single-word adposition or case marker with the meaning of *into* exists in Japanese (i.e., *he* is the best single-word adposition one can get as a translation for *into*). Pairs that do not satisfy either of the criteria (1) and (2) are considered dissimilar.

With these 2 criteria, we can see in Table 4 that the SS-based measure captures more *truly* similar pairs than the CWE-based measure. Although both measures have the same number of pairs satisfying criterion (1) (dictionary translation), the SS-based measure has 10 pairs satisfying criterion (2) (conceptual congruence) as opposed to the CWE-based measure with no such pairs. For example, *except* and *toka* (such as) are opposite in polarity (i.e., exclusion versus inclusion), yet they share the same axis on which the polarity operates (i.e., set membership).

It bears repeating that the CWE-based metric was evaluated as a sort of sanity check; it is *not* surprising that a similarity metric with access to

[17] https://dictionary.cambridge.org/us/dictionary/english-japanese

Metrics	Top 15 EN⇔JP Pairs (Score)					
SS	in place of⇔yori	0.0	**than⇔yori**	0.0	besides⇔toka	0.17
	but⇔toka	0.17	except⇔toka	0.17	nothing but⇔toka	0.17
	besides⇔ya	0.30	but⇔ya	0.30	except⇔ya	0.30
	nothing but⇔ya	0.30	**home⇔he**	0.31	underneath⇔he	0.31
	of⇔no	0.34	**into⇔he**	0.36	at all⇔kurai	0.36
CWE	in spite of⇔nante	0.54	**of⇔no**	0.54	in order to⇔nante	0.52
	in spite of⇔nitsuite	0.52	in spite of⇔kurai	0.52	in order to⇔kurai	0.52
	in order to⇔nitsuite	0.52	in spite of⇔no	0.52	**from⇔kara**	0.51
	in⇔ni	0.51	all over the place⇔nante	0.50	in⇔no	0.50
	away from⇔kara	0.49	in spite of⇔de	0.49	at last⇔nante	0.49

Table 4: Top 15 cross-linguistically similar adpositions and case markers based on SS and CWE metrics. For SS-based metric, lower scores mean higher similarity (smaller divergence), and for CWE-based metric, higher scores mean higher similarity (higher cosine similarity). Rankings read from left to right, row by row. **Boldfaced** cells correspond to dictionary translation, underlined cells correspond to conceptually congruent pairs with differing polarity or specificity, and greyed-out cells correspond to neither.

gold lexical semantic labels should fare better on a lexical semantic evaluation than a metric based on self-supervised fully distributional representations. Still, these results highlight the benefit of the SNACS framework (especially with gold annotations) for cross-lingual analyses.

7. Conclusion

In this paper, we presented the first Japanese SNACS corpus by defining annotation targets, conducting an inter-annotator agreement study, and providing both quantitative and qualitative descriptions of the annotated corpus, with a particular focus on the cases unique to Japanese. We also analyzed cross-lingual similarities between English and Japanese adpositions and case markers via similarity metrics based on supersense distributions, showing that the SS-based metric captures both literal and conceptual similarities, which is not always clear in the CWE-based metric.

8. Acknowledgements

We thank the anonymous reviewers for their constructive and insightful feedback. We also thank Jena D. Hwang for her feedback during the annotation phase of this project. The research was supported in part by NSF award IIS-2144881 (Nathan Schneider, PI).

9. Limitations

It is important to note that the Japanese *Le Petit Prince* is a translated fictional novel, which may contain genre-specific linguistic properties. In other words, the distribution of particles as well as their uses we observed in our data may differ from other Japanese genres/domains. For example, the frequency of the quotative particle と(*to*) may be particularly high due to the abundance of dialogue in the text.

It is also important to note that the results from §6 should be considered a proof of concept. Given the increasing diversity of SNACS-annotated corpora, a more extensive and thorough cross-lingual analyses should be conducted, which was beyond the scope of this paper.

10. Bibliographical References

Nobuo Akiyama and Carol Akiyama. 2012. *Japanese grammar*. Barrons Educational Series.

Aryaman Arora, Nitin Venkateswaran, and Nathan Schneider. 2022. MASALA: Modelling and analysing the semantics of adpositions in linguistic annotation of Hindi. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5696–5704, Marseille, France. European Language Resources Association.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, and Yugo Murawaki. 2018. Universal Dependencies version 2 for Japanese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Austin Blodgett and Nathan Schneider. 2018. Semantic supersenses for English possessives. In *Proc. of LREC*, pages 1529–1534, Miyazaki, Japan.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518*.

Daniel Chen and Mans Hulden. 2022. My case, for an adposition: Lexical polysemy of adpositions and case markers in Finnish and Latin. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2610–2616, Marseille, France. European Language Resources Association.

Naoko Chino. 1991. *All About Particles: A Handbook of Japanese Function Words*. Kodansha International Ltd.

Yasuharu Den, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Minematsu, Kiyotaka Uchimoto, and Hanae Koiso. 2007. The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics. *Japanese linguistics*, 22:101–123.

Jena D. Hwang, Archna Bhatia, Na-Rae Han, Tim O'Gorman, Vivek Srikumar, and Nathan Schneider. 2017. Double trouble: The problem of construal in semantic annotation of adpositions. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 178–188, Vancouver, Canada. 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.

Stefan Kaiser, Yasuko Ichikawa, Noriko Kobayashi, and Hilofumi Yamamoto. 2003. *Japanese: A comprehensive grammar*. Routledge.

Sue A Kawashima. 1999. *A dictionary of Japanese particles*. Kodansha Amer Incorporated.

Michael Kranzlein, Emma Manning, Siyao Peng, Shira Wein, Aryaman Arora, and Nathan Schneider. 2020. PASTRIE: A corpus of prepositions annotated with supersense tags in Reddit international English. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 105–116, Barcelona, Spain. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 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, Sofia, Bulgaria. Association for Computational Linguistics.

Naomi McGloin, M Endo Hudson, Fumiko Nazikian, and Tomomi Kakegawa. 2014. *Modern Japanese Grammar Workbook*. Routledge.

Maitrey Mehta and Vivek Srikumar. 2023. Verifying annotation agreement without multiple experts: A case study with Gujarati SNACS. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10941–10958, Toronto, Canada. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Yu Okubo. 2014. *Anotoki-no oji-kun (Antoine de Saint-Exupery, 1943).* Aozora Bunko.

Keiichiro Okutsu. 1978. *"Boku-wa unagi-da" no bunpou-"da" to "no" (Grammar of "boku-wa unagi-da"-"da" and "no").* Kurosio Publishers, Tokyo.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Jakob Prange and Nathan Schneider. 2021. Draw *mir* a sheep: A supersense-based analysis of German case and adposition semantics. *KI - Künstliche Intelligenz*, 35:291–306.

Eriko Sato. 2021. *Complete Japanese Grammar*. McGraw-Hill.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Archna Bhatia, Na-Rae Han, Tim O'Gorman, Sarah R. Moeller, Omri Abend, Adi Shalev, Austin Blodgett, and Jakob Prange. 2022. Adposition and case supersenses v2.6: Guidelines for English. *arXiv preprint arXiv:1704.02134*.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196, Melbourne, Australia. Association for Computational Linguistics.

Melanie Siegel. 1999. The syntactic processing of particles in Japanese spoken language. In *Proceedings of the 13th Pacific Asia Conference on Language, Information and Computation*, pages 313–320, National Cheng Kung University, Taiwan, R.O.C. National Cheng Kung University, Taiwan, R.O.C.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto. 2016. Universal Dependencies for Japanese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1651–1658, Portorož, Slovenia. European Language Resources Association (ELRA).