JSNACS: Adposition and Case Supersenses for Japanese Joshi

Tatsuya Aoyama[1], Chihiro Taguchi[2], Nathan Schneider[1]

[1]Georgetown University, Department of Linguistics

[2]University of Notre Dame, Department of Computer Science and Engineering

{ta571, nathan.schneider}@georgetown.edu, ctaguchi@nd.edu

LREC-COLING 2024

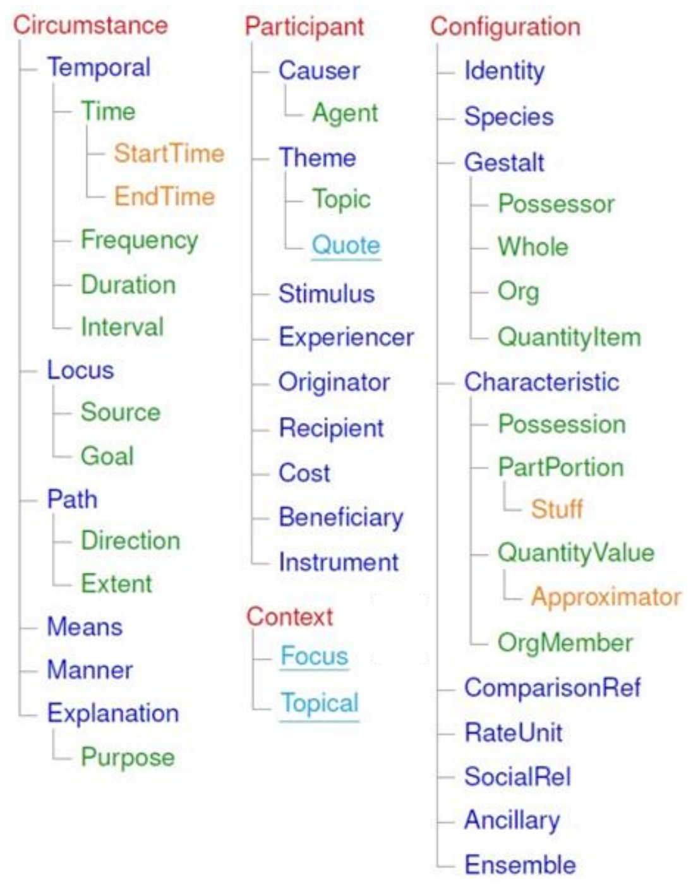
Check out our corpus!

Introduction

- Semantic Network of Adposition and Case Supersenses (SNACS; Schneider et al., 2018) now applied to various typologically different languages.
- Japanese 助詞 (*joshi*), which is often translated as *particles*, do not map to English prepositions in a straightforward manner.
- This study aims at extending SNACS annotation to Japanese.
- Construal Analysis (SceneRole↝Function) (Hwang et al., 2017) and SNACS:

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

(2) It's sad **for**/Experiencer↝Beneficiary Tom.



Research Questions

1. How can we characterize the semantics of Japanese particles using the SNACS framework?
2. Can we use supersense distributions to compare the semantics of adpositions/case markers within and across languages?

Data & Annotation

Japanese translation of *Le Petit Prince (The Little Prince)*, freely available at online[1]

1. The extracted texts were tokenized and UPOS and XPOS tagged fully automatically using MeCab. Segmentation, tokenization, and POS tags were manually corrected where relevant.
2. Supersense was annotated manually by the author in consultation with the original SNACS guideline (Schneider et al., 2020), Korean SNACS guideline (Hwang et al., 2020), and the SNACS website (http://www.xposition.org)

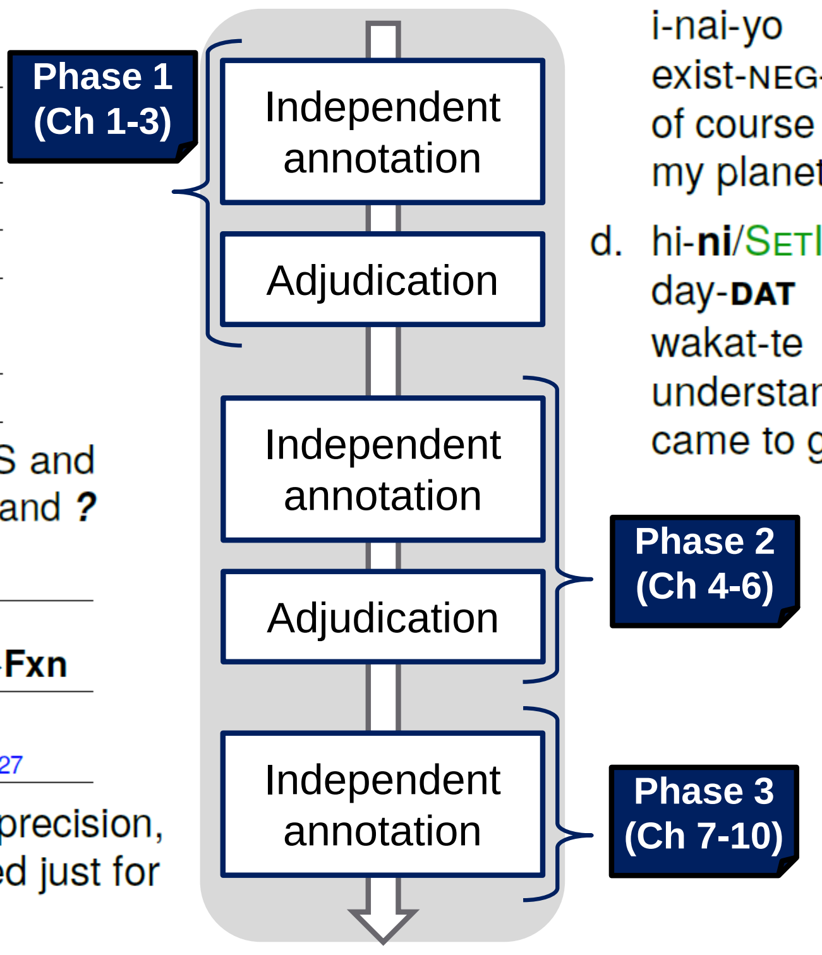
Many-to-many mapping between UPOS and XPOS; particle (binding) was included; particle (adverbial) was included when it *can* modify an NP; particle (nominal), particle (conjunctive), particle (case) that maps to CCONJ, and particle (phrase-final) were all excluded.

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.

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.



Corpus Statistics

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.

Relatively small set of particle types (30) compared to Korean (29) and English (60), and larger set of unique construal types (135) compared to Korean (75) and English (97).

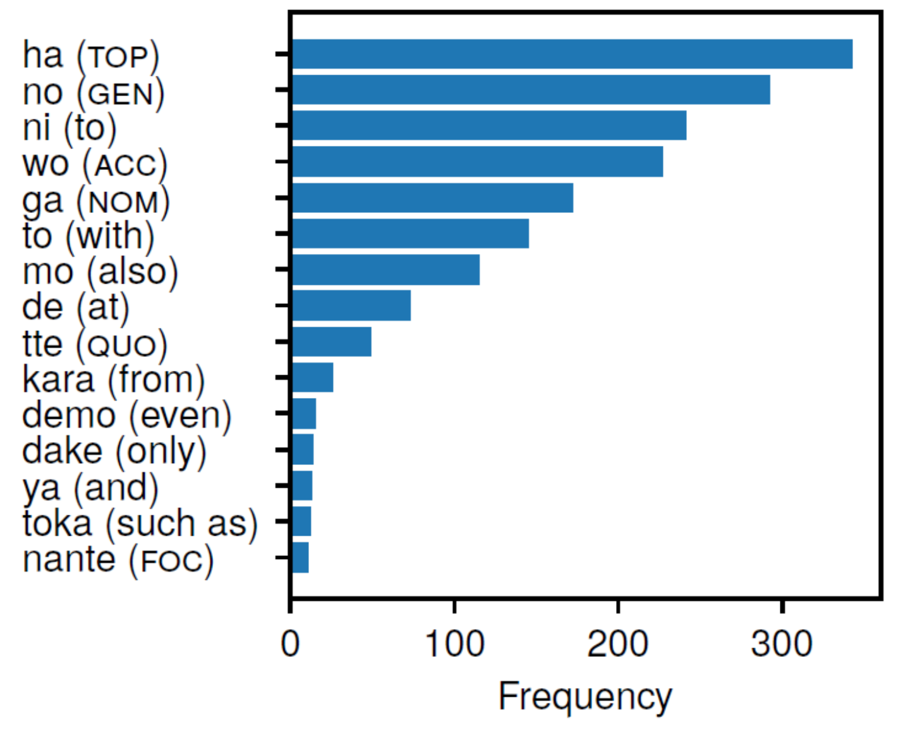


Figure 1: Frequency breakdown by word type o the 15 most common particles.

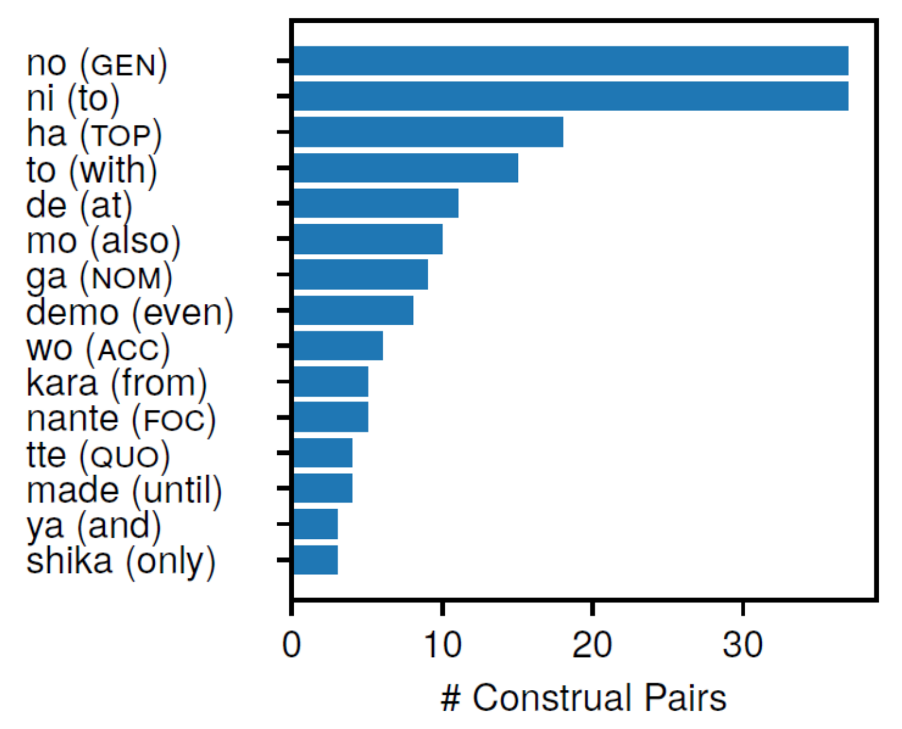


Figure 2: Number of distinct construal pairs for the 15 most polysemous particles.

- Similar to Korean (Hwang et al., 2020), topic marker and case markers (ACC, NOM) are among the most frequent.
- Genitive marker の (-no) and dative particle に (-ni) are among the most polysemous (examples below).

(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*)

(10) a. mottomorashii-**to**/Content↝Quote
likely-**QUO**
omou
think
think that it's likely (*ch4-s34*)

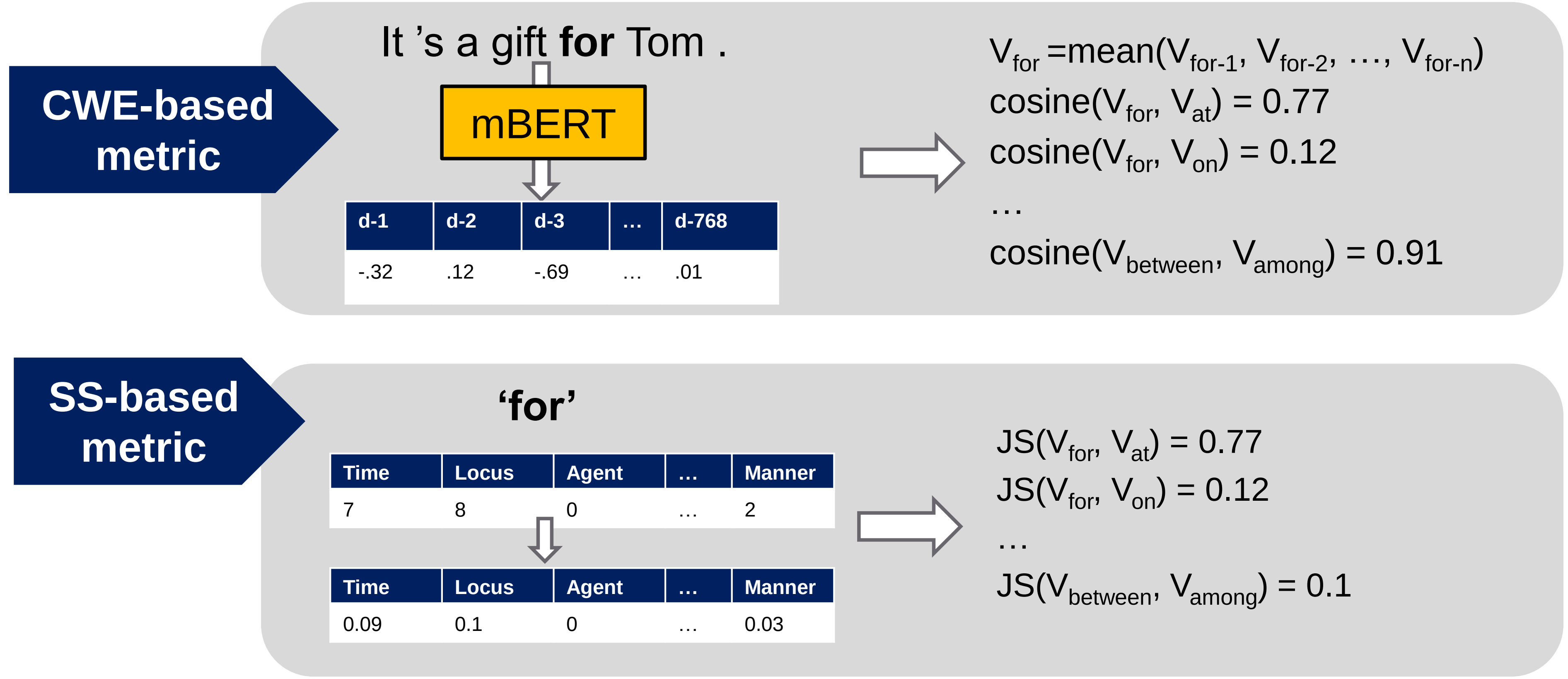
b. tekuteku-**to**/Manner↝Quote isu-wo
trektrek(onomatopoeia)-**QUO** chair-ACC
motte aruke-ba
hold walk-if
if you hold the chair and walk step after step (*ch6-s17*)

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

Polysemy of the particle に (-ni). Also notice that stacking particles is very common in Japanese.

Construal analysis capturing the subtle differences in the usage of the quotative particle と (-to). Contextual meaning is captured in SceneRole and static meaning in Function, making up the construal (SceneRole↝ Function).

Experiments



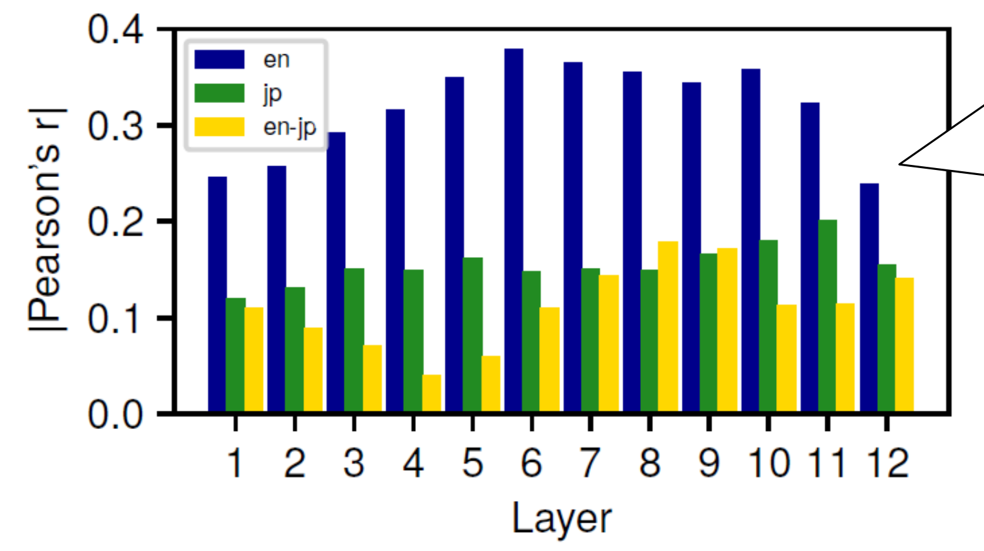
d-1	d-2	d-3	…	d-768
-.32	.12	-.69	…	.01

V_{for} =mean(V_{for-1}, V_{for-2}, …, V_{for-n})
cosine(V_{for}, V_{at}) = 0.77
cosine(V_{for}, V_{on}) = 0.12
…
cosine($V_{between}$, V_{among}) = 0.91

Time	Locus	Agent	…	Manner
7	8	0	…	2

Time	Locus	Agent	…	Manner
0.09	0.1	0	…	0.03

JS(V_{for}, V_{at}) = 0.77
JS(V_{for}, V_{on}) = 0.12
…
JS($V_{between}$, V_{among}) = 0.1



Moderate correlation for within-English setting, weak correlation for within-Japanese setting, and even weaker correlation for cross-lingual setting (English-Japanese). The two metrics seem to be capturing different aspects! Qualitative analyses for comparison:

SS-based metric seems to be capturing what is not captured by the CWE-based metric (fewer greyed-out cells). This is expected, given that SS is manually annotated and CWE is learned in a self-supervised manner for general NLP purposes.

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.

References

- 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.
- 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.
- Nathan Schneider, Jena D Hwang, Archna Bhatia, Vivek Srikumar, Na-Rae Han, Tim O'Gorman, Sarah R Moeller, Omri Abend, Adi Shalev, Austin Blodgett, et al. 2020. Adposition and case supersenses v2. 5: 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.

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