The Propositional Idea Densities of Different Languages in Multi-Lingual Parallel Corpus

Yuka Kaise, Yuto Tsuchiya, and Masanori Oya
Graduate School of Global Japanese Studies, Meiji University
cu245002@meiji.ac.jp, cu245001@meiji.ac.jp
masanori_oya2019@meiji.ac.jp

Abstract

This study reports the propositional idea densities (PIDs) of different languages in parallel corpus in order to investigate whether these densities can function as language-independent measures of syntactic characteristics of sentences. The calculation is based on the Universal Dependencies annotations of dependency types in Parallel Universal Dependencies (PUD), a multi-lingual parallel corpus, and the results show a variety of PIDs across the languages in PUD, which reflect typological variations of information packaging across languages. Some issues of PID also have been pointed out for future research.

1 Introduction

This study reports the propositional idea densities (hereafter, PIDs) of different languages in parallel corpus in order to investigate whether PIDs can function as language-independent measures of syntactic characteristics of sentences. PIDs have been studied using the data of English as a measure of readability and as an indicator of future dementia, yet PID of other languages has not yet been conducted extensively. This study is the first attempt to investigate the PIDs across different languages based on the data of multi-lingual parallel corpus.

2 Previous studies on PID

PID is firmly grounded in well-established psychological theory. Within psycholinguistics, the proposition is considered the basic unit underlying text comprehension and memory (Kintsch & Keenan, 1973). A proposition may comprise diverse linguistic constituents—adjectives, adverbs, verbs, prepositions, and conjunctions—and PID is computed by dividing the number of propositions in a sentence or text by its total word count (Snowdon et al., 1996).

The PID construct has three principal functions: assessing textual readability, forecasting later dementia risk, and gauging sentence complexity in second-language acquisition (SLA) research. With respect to readability, Kintsch and Keenan (1973) showed that passages with lower PID scores are more readily recalled, underscoring the metric's relevance to ease of reading. Extending this work, Covington (2008) compared PID values across genres and observed that introductory and technical documents typically fall below 0.5, whereas research articles display a broader distribution. Such variability in research papers likely stems from their dual role: introducing novel concepts, like introductory texts, while simultaneously conveying detailed technical information, akin to technical documents.

PID has also been investigated as an indicator of future cognitive decline. Empirical evidence indicates that reduced PID in an individual's language output may presage the later emergence of dementia or Alzheimer's disease. In the longitudinal "Nun Study", Snowdon et al. (1996) analyzed autobiographical essays written in early adulthood and found that participants with lower PID scores were more prone to develop Alzheimer's disease five decades later. These results suggest that higher PID scores may signal a preserved capacity to handle syntactically complex structures and could therefore serve as an early marker of cognitive resilience. Similar results were observed in Kemper et al. (2001).

Given its theoretical grounding, PID has been adopted in SLA as an index of sentence complexity and, by extension, learner proficiency (Lopes & Pinto, 2022; Lunn et al., 2022, among others). Differences in learners' PID scores not only reveal variation in the syntactic complexity of their output but may also mirror underlying cognitive capacities for processing complex structures; nonetheless, using PID to predict future neurological decline lies

outside its intended scope.

Notwithstanding its promise, the cross-linguistic study of PID remains sparse. Most existing investigations rely on English data, and systematic analyses of PID in other languages are still lacking. Consequently, our understanding of how PID might operate in linguistic contexts beyond English is limited. Should translations conveying the same meaning exhibit comparable PID values across languages, it would imply that PID functions as a language-independent measure of sentence complexity. Conversely, substantial cross-linguistic divergence in PID for equivalent sentences would suggest that the metric's applicability may be confined to English.

3 This study

3.1 Research questions

This study aims to address the issue explained in the previous section, that is, the lack of cross-linguistic study of PID as a measure of sentence complexity. The research question of this study is as follows:

1. Do sentences of different languages with the same meaning share the same PID?
2. If their PIDs are varied across different languages, what are the cause(s) of the variations of PIDs?

If the answer to the question (1) is affirmative, then PID can be considered as a measure of sentence complexity which can be applied to a variety of languages. If it is negative, then we need to address the question (2) from the viewpoint of typological variations of languages.

3.2 Data

This investigation draws on the Parallel Universal Dependencies Treebanks (PUD) (de Marneffe et al. 2006, 2008; MacDonald et al. 2013; Petrov et al. 2013; Tsarfaty 2013; Zeman 2008; Zeman et al. 2017). Comprehensive documentation of the resource is provided on the CoNLL-2017 shared-task website, "Multilingual Parsing from Raw Text to Universal Dependencies" (http://universaldependencies.org/conll17/).

PUD encompasses 21 languages—Arabic, Czech, Mandarin Chinese, English, Finnish, French, German, Galician, Hindi, Icelandic, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish—each represented by 1,000 sentences that are translations of an identical set of English source texts. The sentences were first morpho-syntactically annotated by Google and subsequently converted to the Universal Dependencies (UD) scheme in accordance with version 2 guidelines by members of the UD community in the CoNLL-U format.

For example, an example sentence "David has been writing several articles on Dependency Grammar and syntactic complexity." is annotated with Universal Dependencies as follows in Table 1 (some annotations have been deleted for the sake of simplicity):

1	David	David	NOUN	4	nsubj
2	has	have	AUX	4	aux
3	been	be	AUX	4	aux
4	writing	write	VERB	0	root
5	several	several	ADJ	6	amod
6	articles	article	NOUN	4	obj
7	on	on	ADP	9	case
8	Dependency	dependency	NOUN	9	compound
9	Grammar	grammar	NOUN	6	nmod
10	and	and	CCONJ	12	cc
11	syntactic	syntactic	NOUN	12	compound
12	complexity	complexity	NOUN	9	conj
13	.	.	PUNCT	4	punct

Table 1: The simplified UD annotation on "David has been writing several articles on Dependency Grammar and syntactic complexity."

Each column contains the following information from the lest to the right: (1) the order of the words in the sentence; (2) the words in the sentence; (3) the lemma (the dictionary form) of these words; (4) the parts of speech of the words; (5) the dependency head of each word; and (6) the dependency type. The first raw reads "The 1st word of this sentence is "David," which has the dictionary form "David," whose part of speech is NOUN; it depends on "writing," the 4th word of this sentence, and its dependency type is *nsubj* (nominal subject).

One of the characteristics of UD is that it focuses on the dependencies among content words and function words are all dependent on content words. For example, auxiliaries are dependent on the verbs which they add modal meanings, and prepositions are dependent on the nouns which follow them. In the example above, "has" and "been" are dependent on "writing" with the dependency type "aux," and "on" is dependent on "Grammar" with the dependency type "case." UD has chosen this annotation policy based on the insight that the meaning expressed by function words in certain languages (e.g., English) can be expressed not by

	ar	ch	cz	de	en	es	fi	fr	gl	hi	id
acl	34	20	112	18	193	116	223	150	429	338	246
acl:relcl	320	448	239	271	211	244	227	226	0	215	511
advcl	316	516	189	223	292	180	283	218	211	200	369
advmod	448	1225	661	1124	845	823	872	865	804	381	971
amod	1620	419	1817	1100	1348	1311	910	1394	1286	1412	585
ccomp	287	403	172	169	135	148	167	174	184	153	97
compound	386	1777	21	369	864	209	181	0	23	1277	35
conj	661	383	731	841	635	656	688	651	653	600	664
csubj	57	72	57	28	27	41	2	23	32	0	25
case	3047	1665	1857	2055	2511	3696	318	3208	3652	4076	1865
nummod	150	809	319	227	195	191	312	218	310	279	359
parataxis	24	3	23	68	97	105	108	107	90	94	114
root	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
xcomp	152	537	248	190	271	470	159	396	263	584	225
sum	8502	9277	7446	7683	8624	9190	5450	8630	8937	10609	7066
all	20747	21415	18463	21332	21126	23751	15813	24369	23309	23725	19858
PID	0.410	0.433	0.403	0.360	0.408	0.387	0.345	0.354	0.383	0.447	0.356

	is	it	ja	kr	pl	pt	ru	sv	th	tr
acl	171	208	1100	0	249	180	256	132	976	515
acl:relcl	303	241	0	1188	179	233	160	301	613	0
advcl	235	250	916	999	176	119	197	341	341	455
advmod	847	777	314	593	484	768	909	887	1190	574
amod	881	1395	84	208	1423	1328	1791	1253	654	1318
ccomp	124	137	75	68	85	119	131	122	275	173
compound	174	48	3061	2359	0	20	9	263	1927	519
conj	746	662	549	409	711	647	695	658	662	696
csubj	43	34	8	19	5	29	48	35	49	92
case	2132	3443	6496	404	1996	3604	2121	2225	2413	692
nummod	269	202	432	487	86	201	183	275	372	268
parataxis	85	99	0	0	0	102	195	134	5	15
root	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
xcomp	288	249	0	0	198	387	331	230	1070	127
sum	7298	8745	14035	7734	6592	8737	8026	7856	11547	6444
all	18835	23570	28788	16488	18384	23277	19355	19076	22289	16720
PID	0.387	0.371	0.488	0.469	0.359	0.375	0.415	0.412	0.518	0.385

Table 2: The frequencies of propositional dependency types and PIDs of the 21 languages in PUD. Abbreviations: ar: Arabic; ch: Chinese; cz: Czech; de: German; en: English; es: Spanish; fi: Finnish; fr: French; gl: Galician; hi: Hindi; id: Indonesian; is: Icelandic; it: Italian; ja: Japanese; kr: Korean; pl: Polish; pt: Portuguese; ru: Russian; sv: Swedish; th: Thai; and tr: Turkish.

independent words but by morphemes in content words in other languages (e.g., Russian). Focus on the dependencies among content words allows UD to capture cross-linguistic parallelism of the dependencies among them.

Because the dataset consists of semantically aligned translation pairs, cross-linguistic syntactic variation—including differences in dependency distances—can be analyzed while holding meaning constant.

3.3 Method

PIDs of the sentences in PUD are calculated based on the distinctions between propositions and non-propositions according to the type of dependency with which each word in a sentence depends on another in the same sentence. Propositions in this study are those words that depend on other words with the following 14 dependency types: *acl* for the verbs in adjectival participle clauses, *acl_relcl* for the verbs in relative clauses, *advcl* for the verbs in adverbial clauses, *advmod* for adverbs, *amod* for adjectives, *ccomp* for the verbs of clausal complements, *compound* for nominal compounds, *conj* for conjunctions, *csubj* for clausal subjects, *case* for prepositions, *nummod* for modifications by numericals, *parataxis* for paratactic phrases, *root* for the main verb of a clause, and *xcomp* for the verbs in external complements, which are those whose

subjects are either the subject or the object of the main clause. Words with these dependency types are expected to cover those defined as propositions according to Snowdon et al. (1996), which are adjectives, adverbs, verbs, and conjunctions. For each of the 21 languages in PUD, the number of these dependency types are calculated, then it is divided by the sum of the dependencies to obtain the PID of the language.

3.4 Results

Table 2 summarizes the result of calculating PIDs of the 21 languages in PUD. The mean of the PIDs of these 21 languages is .403, and their SD is .046. Approximately 40% of all tokens in PUD encode propositional content. The correlation between the total word counts and the PIDs of these languages is weak (r = 0.266). The top 3 PIDs are Thai, Japanese, and Korean, and the bottom 3 are Finnish, French, and Indonesian. The most cross-linguistically frequent dependency type is *case* (mean: 2546.48), which is followed by *amod* and *advmod*. The most cross-linguistically variable type is also *case* (SD: 1393.27), followed by *compound* and *amod*.

These frequent dependency types seem to have language-specific influence on the PID of a given language. For example, *case* is about 46% of the propositions in Japanese, while it is about 6% of those in Finnish; *compound* is about 21% in Japanese, while about 3% in Finnish, and 0% in French.

3.5 Discussions

PID reflects the density of propositional content encoded in a language. A higher PID of a language may indicate that it compresses more semantic content into more propositions (verbs, adjectives, adverbs, or conjuncts), which can show its structural compactness. Specifically, languages with frequent use of prepositions or postpositions (e.g., Thai and Japanese) show higher PIDs, while languages in which grammatical relations are expressed by inflections (e.g., Finnish, Turkish) show lower PIDs. This suggests that PIDs of different languages indirectly capture their typological variation. As such, PID allows us to measure how different languages distribute syntactic and semantic load. This makes PID a practical tool for comparing the surface density of propositional content across languages with different morphosyntactic strategies.

We need to point out some limitations on PIDs as a measure for syntactic characteristics of sentences: First, as the name PID indicates, it is a measure of density, not structural depth or complexity per se. This means that a high PID does not necessarily indicate a more complex syntax; rather, it may simply reflect fewer function words or more compact morphosyntax.

Second, PID should not be considered as a measure of sentence *complexity*, because it does not necessarily focus on the embeddedness of the syntactic structure, which is one of the factors of syntactic complexity. It is true that a sentence has a larger number of propositions if it contains many adverbial clauses and relative clauses (hence, more embedded) because these clauses contain verbs and possibly more adjectives, adverbs and prepositions, yet this also means it contains a larger total word count, hence its PID does not increase.

Third, the issue of annotation bias must be addressed across a variety of languages. PID is heavily influenced by the segmentation of sentence strings into words and annotation of them with parts of speech tags. The results that Japanese or Thai appear denser than others in PUD may be due to the annotation policy that postpositions or compound markers are counted as separate tokens, and we can expect that parallel corpora with different annotation policies may yield different results of PIDs for them.

4 Conclusion

This study reported the propositional idea densities (PIDs) of different languages in parallel corpus in order to investigate whether these densities can function as language-independent measures of syntactic characteristics of sentences. The calculation is based on the Universal Dependencies annotations of dependency types in Parallel Universal Dependencies, a multi-lingual parallel corpus, and the results show a variety of PIDs, which reflect typological variations of information packaging across languages. Three issues (PID not as a syntactic complexity measure, lack of consideration on the embeddedness, and annotation biases) have been raised for future research on propositional idea densities for characterizing syntactic properties of sentences in natural languages.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 24K04089.

References

Michael Covington. 2009. Idea Density — A Potentially Informative Characteristic of Retrieved Documents. *IEEE Southeastcon 2009*. https://ai1.ai.uga.edu/caspr/Covington-2009-Idea-Density-paper-SEC09-060.pdf

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. *Proceedings of LREC*.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.

Susan Kemper, Lydia Greiner, Jane Marquis, Katherine Prenovost and Tracy L Mitzner. 2001. Language decline across the life span: Findings from the nun study. *Psychology and Aging*, 16(2), 227–239. https://doi.org/10.1037/0882-7974.16.2.227

Walter Kintsch and Janice Keenan. 1973. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 5, 257–274.

Ângela Filipe Lopes and Maria da Graça Lisboa Castro Pinto. 2022. Assessing L2 Portuguese writing: idea density and sentence complexity. *Signo*, 47(88), 73–86.

Andrew M Lunn, Daniel Matthias Bürkle, Rebecca Ward, Alice P McCloskey, Adam Rathbone, Aaron Courtenay, Rachel Mullen, Andrea Manfrin. 2021. Spoken propositional idea density, a measure to help second language English speaking students: A multicentre cohort study. *Medical Teacher*, 44(3), 267–275. DOI: 10.1080/0142159X.2021.1985097

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. *Proceedings of ACL*.

Marije C Michel, Folkert Kuiken and Ineke Vedder. 2007. The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, 45, 241–259.

James R. Miller and Walter Kintsch. 1980. Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 335–354.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. *Proceedings of LREC*.

David A. Snowdon, Susan Kemper, James Mortimer, Lydia Greiner, David R. Wekstein, and William R. Markesbery. 1996. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *JAMA*, 275, 528–532.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of Stanford dependencies. *Proceedings of ACL*.

Daniel Zeman. 2008. Reusable Tagset Conversion Using Tagset Drivers. *Proceedings of LREC*.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver.