Where Is the Head Positioned in Indonesian Language?: A Corpus Study of Head Directionality from a Dependency Perspective

***Abstract*—The link connecting two constituents in a syntactic dependency relation has been a key measure of interest in the studies of syntax. This include the positioning of a *head* word in regards to its *dependent* or the head directionality. The direction of the link connecting both constituents is a familiar topic in syntax through a phrase structure perspective. However, it is yet to be explored using dependency approach, particularly in Indonesian language. This study uses a large-scale corpus of Indonesian language in written and spoken manner. The preliminary findings exhibit that there are particular conditions in which head-initial dependency direction is preferred over head-final relation. The analysis also indicates a preference to position the head before its dependent albeit producing a longer dependency distance. Overall evaluation shows that aside to the minimization of dependency distance, Indonesian speakers might choose a certain position of the head to better facilitate cognitive demands and reduce the burden bear by the working memory.**

***Keywords*-dependency; head directionality; syntax; Indonesian; dependency direction**

I. Introduction

There are several concepts that relate syntactic structures of a sentence with human's cognition, such as dependency. In a *dependency* theory, every linguistic unit is connected with each other by a link that represents direct asymmetrical syntactic relation forming an iterative structure attaching all the constituents within a sentence [8]. This syntax framework contributes greatly to the discussion on how far a language development is shaped by the mechanical constraints of human cognition, particularly in relation to language acquisition and language application [1]. Heringer proposed a linear distance that represents dependency between two words or constituents, which has been adopted into modern computational methods of quantifying dependency [2]. This linear distance is measured by counting the constituents involved between the connected constituents. However, another important aspect in measuring dependency, in addition to the linear distance, is the position of the *head* in regards to its *dependent*, or the *head directionality*.

A. Measuring dependency

Words are the main constituents of syntax that carry with them the grammatical relation that connect one word with another [8]. The term *dependency* itself is a derivative of a connection that refers to an asymmetrical relation between a superordinate and its subordinate [**?**], [2], [8]. Many studies and theory developments have all described in detail what a dependency structure between two constituents looks like [8], [**?**], [**?**]. It can be summarized from these references that a dependency structure consists of a

Figure 1. Elements of a dependency relation

binary and asymmetrical relation between two linguistic units and the type of a dependency relation is usually indicated using a label on top of an arc linking the two constituents, as shown in 1. An arrow pointing from a more superior constituent to a more inferior one represents the hierarchical relation between two constituents. The governor is also more commonly called as a *head* and the subordinate is generally called a *dependent*. In a more complex and recursive structure of dependency relations, such as a sentence, the sentence head is usually called a *root*.

The arc in 1 illustrates how a constituent must be kept active in working memory until both constituents are realized in an utterance, which result in the integration of meanings carried by both constituents [**?**], [**?**]. Therefore, the greater the dependency distance, the heavier the burden borne by the working memory. To measure a dependency distance (DD), a sentence is treated as a string of constituents $C_1...C_i...C_n$. Each constituent has a significant index that corresponds its position within the given sentence. For example, C_1 indicates that the constituent is located on the first position in the sentence. There are different approaches of measuring dependency distance. Largely, they involve the value of DD, which is obtained by subtracting directly linked constituents [**?**], [**?**], [4].

Previous substantial studies on measuring dependency used dependency length (DL) as their main approach [3], [4]. DL is the sum of absolute DD of an entire sentence, which can be defined as follows:

$$\sum_{i=1}^{n-1} |DD_i|$$

While other notable studies used a different approach called the mean dependency distance (MDD), which divides DL with the number of dependency relations of an entire sentence [**?**], [**?**]. MDD can be defined as follows:

$$\frac{1}{n-1} \sum_{i=1}^{n-1} |DD_i|$$

Figure 2. Same-branching dependency direction [6]

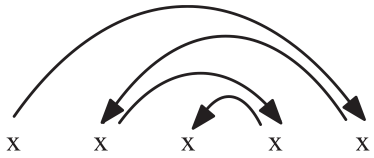

Figure 3. Mixed-branching dependency direction [6]

B. Head directionality

There has yet a convention on head directionality in regards to the online processing in working memory, particularly for Indonesian language. However, previous attempts have been done to provide details on the head's position and its relation to grammars. Several studies in the early development of dependency theory found that a language tends to apply a consistent dependency direction, whether it is *head-first*/*head-initial* or *head-last*/*head-final* [**?**], [**?**], [**?**]. Hawkins and Frazier assumed that this consistency, as shown in 2, is the implementation of a strategy to minimize the distance between a head and its dependent [**?**], [**?**].

Temperley argues that a consistent dependency direction or *same-branching* is not an ideal representation of how head directionality works in real utterances [6]. This argument supports Dryer's research who found that the conventional view of same branching only applies on phrases consist of many constituents, while the dependency direction of phrases with single or less words are inconsistent [7]. Gildea and Temperley mentioned this research and provided empirical evidence that exhibit a combination of same branching and mixed branching (3) dependency direction to produce a more balanced head-initial and head-final mixture to yield the shortest dependency distances (4) [3].

II. Methods

The written data, obtained from a partnership with an Indonesian technology company called Dattabot, consists of 9311 sentences of various news articles published from 2008 to 2018. The spoken data, obtained from partnerships with various journalists across the country, consists of 10.219 sentences of live reports, interviews, and other spontaneous utterances (all in the journalistic domain) recorded from 2010 to 2018. The main challenge in including spoken data is because computational parsers are mostly trained to parse written data. Therefore, sentences with specific features of spoken language, such as hesitations, fillers, etc., are not included in the corpus. These two sets of corpora are preprocessed, parsed, and cleaned using UDPipe [**?**] and manual verifications. UDPipe is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files based on dependency treebanks provided by Universal Dependencies 2.0 that are contextually adjusted for many languages, including Indonesian [**?**], [**?**]. This study uses third party CRAN package UDPipe (version 0.5) on R programming language (version 3.3.3) [**?**], [**?**].

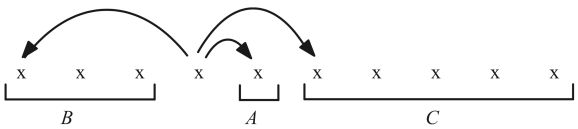

Figure 4. Head-directionality balance [6]

The parsed data are annotated based on the parsing index to measure the dependency distances, length, and mean dependency distances. Ini the parsing index, $root$ has an index of 0. The indexing for the remaining words starts from 1. As mentioned above on Heringer's approach of measuring dependency, this study also uses the number of constituents involved in a dependency relation between two directly-linked constituents as a measuring unit. Therefore, adjacent and directly-linked constituents produce a dependency distance of 1. Data classification, subsetting, and dependency measures are performed on R and Python programming language version 3.3.3 and 3.6.3 respectively [**?**], [**?**]. As visualized in 1, a dependency relation is illustrated through an arc with an arrow. The arrow or direction represents the hierarchy between linked constituents. For Indonesian language, an arrow pointing to the right illustrates a positive $(+)$ dependency relation and denotes a head-initial dependency direction. On the other hand, an arrow pointing to the left illustrates a negative $(-)$ dependency relation and denotes a head-final dependency direction.

A central node in a sentence suggests the main argument or relations of the sentence based on dependency. In a more complex sentence, a branch in a central node can contain multiple constituents. For this reason, the head directionality analysis is performed on two types of relation: (i) on directly-linked constituents in all levels of a dependency tree, including central nodes and adjacent relations, and (ii) on central nodes only, assuming all constituents under each branch are represented by the main arguments. Considering the possible influence sentence length may have on dependency distances, this study classifies both spoken and written data into five categories to acquire more accurate measures [**?**], [**?**], [**?**], [**?**]:

- Sentences with 5 constituents or less.
- Sentences with 6 to 10 constituents.
- Sentences with 11 to 20 constituents.
- Sentences with 21 to 30 constituents.
- Sentences with more than 30 constituents.

III. Results

Indonesian language is considered a free-word order language [**?**]. However, Sneddon describes the many word order rules that put modifiers after their heads based on the functional perspective of phrase structure [**?**]. This study

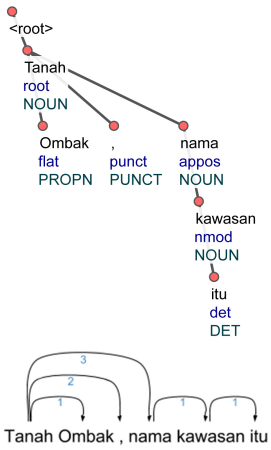

Figure 5. Sentence with all head-initial dependency directions

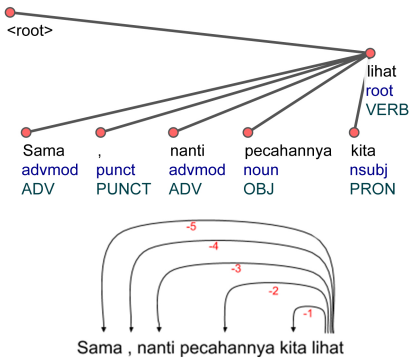

Figure 6. Sentence with all head-final dependency directions

considers this functional view and uses it as a basis to see whether the tendency applies in a dependency structure. This analysis sets apart the head-first (positive) with the head-final (negative) dependency directions to capture a comprehensive characteristic based on the sentence length classifications.

A. Head-initial preference on all levels of a dependency structure

This part of analysis measures dependency distance and number of appearance between head-initial and head-final dependency directions on all levels of a dependency structure. A full consistency of direction such as shown in 5 and 6 are found only in shorter sentences. 5 illustrates a sentence with all head-initial dependency directions, in contrast to 6. However, the number of appearance between two speech modes is too few to be considered as a preference.

As shown in 7 and 8, both speech modes show preference towards the head-initial dependency direction. The ration difference is increasingly larger in longer sentences. There is even an indication of a threshold to reduce the distance of head-final dependency in the spoken data 8. As seen in table I and II, there are more sentences with longer head-initial dependency length and more head-initial dependencies starting from 6 constituents. The

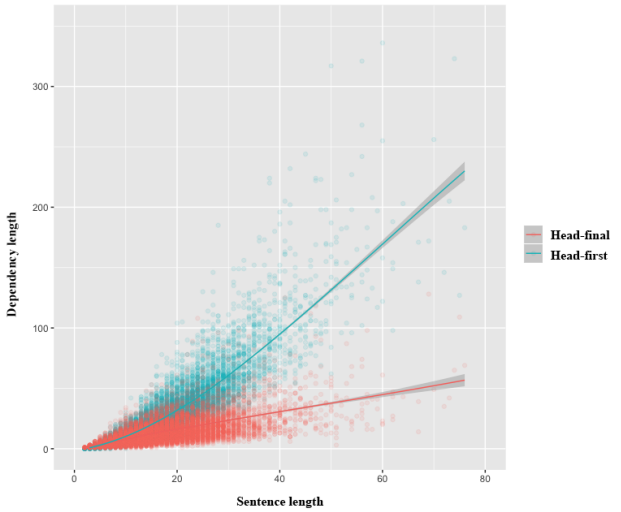

Figure 7. Comparison of head-first and head-final dependencies on all levels based on its length in written data

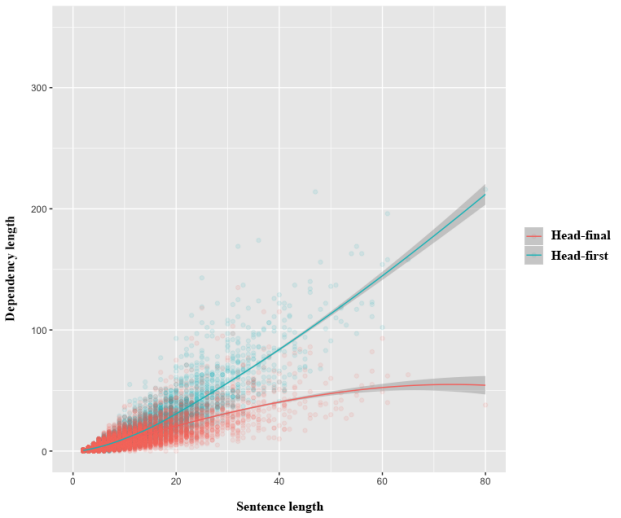

Figure 8. Comparison of head-first and head-final dependencies on all levels based on its length in written data

difference between both directions are increasingly larger particularly in longer sentences. However, the margins of difference are significantly larger in the written data.

Based on descriptive statistics using both methods of dependency length (III) and mean dependency distance (IV), preference for head-initial dependencies occur mostly in adjacent relations. This finding support the application of grammar as mentioned by Sneddon who describes that many word order rules put modifiers after their heads based on the functional perspective of phrase structure [**?**]. Both methods show that head-final dependencies are relatively shorter than head-first dependencies in longer sentences. While in shorter sentences, the length and distance between both directions are quite similar.

Table I
NUMBER OF OCCURRENCE WHERE ONE DEPENDENCY DIRECTION IS MORE DOMINANT THAN THE OTHER ON ALL LEVELS OF A DEPENDENCY STRUCTURE

Length	Written (+ >-)	Written (+ <-)	Spoken (+ >-)	Spoken (+ <-)
<= 5	164	221	1188	1354
6 - 10	992	646	1469	1349
11 - 20	3145	952	1858	1014
21 - 30	1854	200	644	166
31 - 40	558	36	204	37
>40	201	5	89	7

Table II
NUMBER OF OCCURRENCE FOR EACH DEPENDENCY DIRECTION ON ALL LEVELS OF A DEPENDENCY STRUCTURE

Length	Written (+)	Written (-)	Spoken (+)	Spoken (-)
<= 5	677	700	2841	2764
6 - 10	7023	5715	7649	6918
11 - 20	35597	24426	15554	12898
21 - 30	31036	18163	7564	5939
31 - 40	12904	6968	3232	2414
>40	6583	3240	1928	1426

B. Minimization of head-final dependency distance in longer sentences on central node level

The second part of analysis measures dependency distance and number of occurrence between head-initial and head-final dependencies on central node level. 9 and 10 are two sentences found in the written data that have the same dependent clause *kalau boleh tahu* or "if i may know" in English. Even though each sentence has a different position for this dependent clause, there is no difference in terms of their meanings. The similar example also mentioned by Sneddon in the Indonesian Reference Grammar as one of the key feature of Indonesian as a free-word order language [**?**].

Table III
DEPENDENCY DISTANCES OF ALL LINKED CONSTITUENTS ON ALL LEVELS OF A DEPENDENCY STRUCTURE

	+				-				
Length	min	med	max	mean	min	med	max	mean	
<= 5	1	1	4	1,409	1	1	4	1,606	Written
6 - 10	1	1	9	1,866	1	1	9	1,948	
11 - 20	1	1	19	2,468	1	1	19	2,147	
21 - 30	1	1	29	2,971	1	1	29	2,241	
31 - 40	1	1	39	3,597	1	1	37	2,299	
>40	1	1	68	4,22	1	1	55	2,282	
<= 5	1	1	4	1,466	1	1	4	1,534	Spoken
6 - 10	1	1	9	1,933	1	1	9	2,071	
11 - 20	1	2	19	2,535	1	1	19	2,321	
21 - 30	1	2	28	3,263	1	1	27	2,5	
31 - 40	1	2	37	3,718	1	1	32	2,657	
>40	1	2	66	4,067	1	1	46	2,438	

Table IV
MEAN DEPENDENCY DISTANCES OF ALL SENTENCES ON ALL LEVELS OF A DEPENDENCY STRUCTURE

	+				-				
Length	min	med	max	mean	min	med	max	mean	
<= 5	1	1	3,5	1,372	1	1,5	4	1,545	6*Tulis
6 - 10	1	1,667	5,2	1,839	1	1,667	9	1,895	
11 - 20	1	2,25	7,25	2,432	1	1,875	10	2,117	
21 - 30	1	3,765	10,278	2,97	1	2	12,75	2,212	
31 - 40	1,458	3,368	9,933	3,595	1	2	8	2,272	
>40	1,929	3,756	10,567	4,154	1	2	7,111	2,258	
<= 5	1	1	4	1,378	1	1,33	4	1,456	6*Lisan
6 - 10	1	1,714	5,6	1,867	1	1,8	9	2,003	
11 - 20	1	2,286	9,625	2,479	1	2	6,818	2,255	
21 - 30	1,4	3	8,412	3,252	1	2,188	8,636	2,447	
31 - 40	1,556	3,512	7,25	3,719	1	2,27	8,803	2,612	
>40	2,174	3,865	6,895	3,957	1,059	2,294	4,2	2,363	

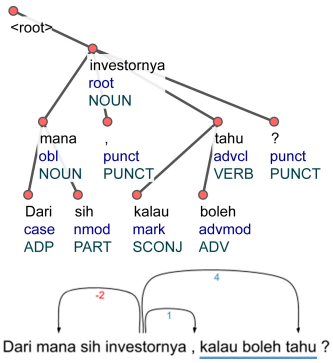

Figure 9. Sentence with a final dependent clause

In 9, all of the constituents forming the clause *kalau boleh tahu* collectively creating a head-initial dependency against its head *investornya* or "the investor" in English. On the other hand, the dependent clause in 10 creates a head-final dependency against its head *ordernya* or "the order" in English. In a dependency theory, both constituents forming a dependency must be stored in the working memory until both are realized [8]. In this head-final dependency case, the constituents in this dependent clause (or branch of a dependency structure) must be stored as a whole until the head is realized.

Central node level comprises of constituents that construct the main argument of a sentence. This also relates to how the root is able to bind other constituents, or its valency. To produce a more accurate overview of a main argument in a sentence, this study focuses the analysis on central node level with verbal roots. Sentences with verbal roots are found around 84,61% or 7874 sentences in written data and 70,91% or 7239 sentences in spoken data, which makes it the most common root type in the data.

As opposed to the analysis on all levels of a dependency structure, the preference to either dependency direction is not as clear, albeit a visible preference for head-initial relation in longer sentences as seen in 11 and 12. There is also an indication of a threshold to reduce the distance of head-final dependency in the spoken data 12. V shows that sentences with more head-initial relations are found in longer sentences (starting 21 constituents). Particularly in spoken data, the number of occurrences between both dependency directions is almost balanced on central node level. This evidence supports Dryer's initial argument that natural language will exhibit a more balanced head-initial and head-final mixture to produce the shortest dependency distances [7].

Despite the balance in the number of occurrence between both dependency directions, VII shows a consistency in result with III. Both analysis (on central node or all dependency levels) show that there are preference to minimize the dependency distance of a head-final relation. This preference can be seen from sentence length 11 constituents, particularly in longer sentences.

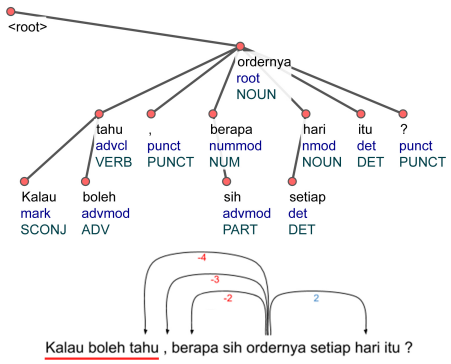

Figure 10. Sentence with an initial dependent clause

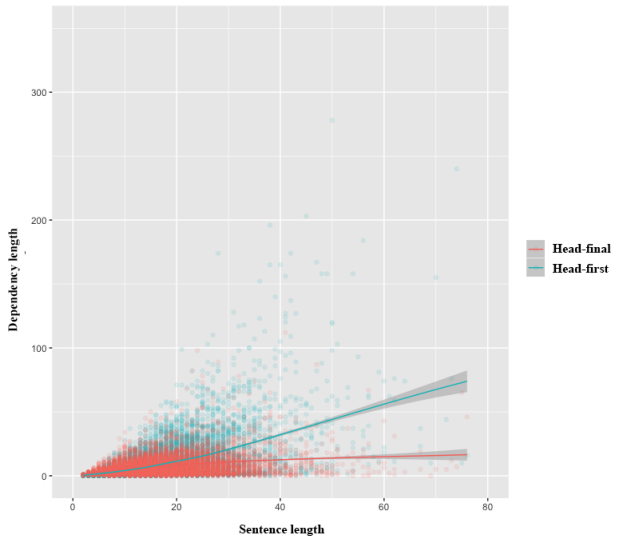

Figure 11. Comparison of head-first and head-final dependencies on central node level in written data

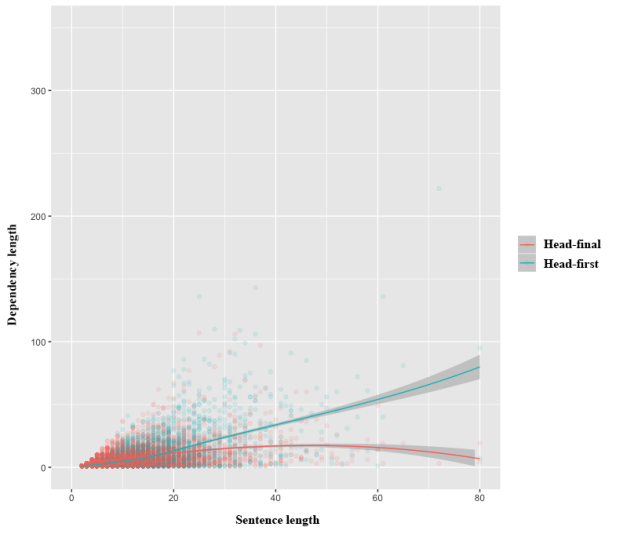

Figure 12. Comparison of head-first and head-final dependencies on central node level in spoken data

Table V
NUMBER OF OCCURRENCE WHERE ONE DEPENDENCY DIRECTION IS MORE DOMINANT THAN THE OTHER ON CENTRAL NODE LEVEL FOR VERBAL *root*

Length	Written (+ >-)	Written (+ <-)	Spoken (+ >-)	Spoken (+ <-)
<= 5	57	180	393	1008
6 - 10	475	796	655	1356
11 - 20	1638	1804	964	1336
21 - 30	1028	778	384	291
31 - 40	325	177	121	85
>40	137	37	61	30

Table VI
NUMBER OF OCCURRENCE FOR EACH DEPENDENCY DIRECTION ON CENTRAL NODE LEVEL FOR VERBAL *root*

Length	Written (+)	Written (-)	Spoken (+)	Spoken (-)
<= 5	201	412	1204	2151
6 - 10	1917	2532	2998	4433
11 - 20	6914	7144	4753	5570
21 - 30	4455	3691	1869	1754
31 - 40	1528	1031	540	461
>40	625	321	315	258

IV. CONCLUSION

Based on the phrase structure in the current discussion of Indonesian grammar, Indonesian is considered a head-initial language, where the governor of both constituents are positioned before the dependent [**?**], [**?**]. However, there has yet empirical evidence to verify this assumption using a large-scale performance data, particularly for spoken language. Even though no convention has been made whether head-initial languages produces shorter dependency distance compared to head-final language, a large-scale research based on 37 languages has shown empirical evidence that Indonesian language has a higher degree of minimization compared to the other languages,

Table VII
DEPENDENCY DISTANCES OF ALL LINKED CONSTITUENTS ON CENTRAL NODE LEVEL FOR VERBAL *root*

	+				-				
Length	min	med	max	mean	min	med	max	mean	
<= 5	1	1	3	1,428	1	1	4	1,748	Written
6 - 10	1	2	9	2,336	1	2	9	2,571	
11 - 20	1	3	18	4,083	1	2	19	3,61	
21 - 30	1	4	27	6,265	1	3	29	4,793	
31 - 40	1	6	38	9,046	1	3	37	5,925	
>40	1	9	68	12,88	1	3	44	6,791	
<= 5	1	1	4	1,482	1	1	4	1,69	Spoken
6 - 10	1	2	9	2,28	1	2	9	2,618	
11 - 20	1	3	19	3,826	1	2	18	3,658	
21 - 30	1	4	26	6,413	1	3	27	4,789	
31 - 40	1	6	35	8,638	1	4	32	6,505	
>40	1	7	66	10,27	1	3	46	5,852	

Table VIII
MEAN DEPENDENCY DISTANCES OF ALL SENTENCES ON CENTRAL NODE LEVEL FOR VERBAL *root*

	+				-				
Length	min	med	max	mean	min	med	max	mean	
<= 5	1	1	3	1,363	1	1,5	4	1,632	Written
6 - 10	1	2	7	2,091	1	2	9	2,385	
11 - 20	1	3	16	3,545	1	2,75	19	3,36	
21 - 30	1	5	19,5	5,514	1	3,333	28	4,442	
31 - 40	1	7,5	26	8,044	1	4	36	5,448	
>40	1	11,1	34,286	11,84	1	4,2	42	6,264	
<= 5	1	1	3	1,327	1	1,5	4	1,531	Spoken
6 - 10	1	1,667	5,6	1,801	1	2	9	2,603	
11 - 20	1	2,286	7,167	2,479	1	2,125	6,818	2,315	
21 - 30	1,4	3,118	8,412	3,375	1	2,182	8,636	2,460	
31 - 40	1,947	3,643	7,25	3,879	1	2,353	7,5	2,688	
>40	2,147	4,051	8,697	4,156	1,059	2,357	5,053	2,47	

particularly those with head-final preference [4].

This study showcases evidence for head-directionality analysis on both level: (i) between two constituents on all dependency levels, and (ii) on central node level. By comparing both findings, this study found significant difference related to the preference for a certain dependency direction. Based on both written and spoken data, there is a preference in Indonesian language to minimize the dependency distance of a head-final relation. This preference is visible as early as 11 constituents, and more evident in longer sentences.

However, when looking at the number of occurrence where one dependency direction is more dominant than the other, both findings show obvious difference. On the relation between two constituents on all dependency levels, there is a preference for head-initial relation as early as 6 constituents. While on central node level, both speech modes exhibit a more balanced combination between head-initial and head-final dependency directions, particularly from 11 to 30 constituents. This supports Gildea and Temperley's arguments that a more balanced combination of both dependency directions better illustrates the preference in a natural language [3].

REFERENCES

[1] J. B. Plotkin and M. A. Nowak, "Language evolution and information theory," *Journal of Theoretical Biology*, vol. 205, no. 1, pp. 147–159, 2000.

[2] H. J. Heringer, "Dependency syntax-basic ideas and the classical model," *Syntax-An International Handbook of Contemporary Research*, vol. 1, pp. 298–316, 1993.

[3] D. Gildea and D. Temperley, "Do grammars minimize dependency length?" *Cognitive Science*, vol. 34, no. 2, pp. 286–310, 2010.

[4] R. Futrell, K. Mahowald, and E. Gibson, "Large-scale evidence of dependency length minimization in 37 languages," *Proceedings of the National Academy of Sciences*, vol. 112, no. 33, pp. 10 336–10 341, 2015.

[5] H. Liu, "Dependency direction as a means of word-order typology: A method based on dependency treebanks," *Lingua*, vol. 120, no. 6, pp. 1567–1578, 2010.

[6] D. Temperley, "Dependency-length minimization in natural and artificial language," *Journal of Quantitative Linguistics*, vol. 15, no. 3, pp. 256–282, 2008.

[7] M. S. Dryer, "The Greenbergian word order correlations," *Language*, vol. 68, pp. 81–138, 1992.

[8] L. Tesnière, *Éléments de syntaxe structurale*. Paris, France: Librairie C. Klincksieck, 1959.

[9] G. K. Zipf, *Human behaviour and the principle of least-effort*. Cambridge, Massachusetts: Addison-Wesley Press, 1949.