SYNTACTIC COMPLEXITY COMBINING DEPENDENCY LENGTH AND DEPENDENCY FLUX WEIGHT

CHUNXIAO YAN[*1], SYLVAIN KAHANE[1]

[*]Corresponding Author: yanchunxiao@yahoo.fr
[1]MoDyCo, Université Paris Nanterre & CNRS, Paris, France

1. Introduction

According to psychological research by Miller (1956), human memory is constrained to 7±2 elements. This cognitive constraint has an influence on natural languages at the level of syntax. Yngve (1960) proposed the notion of depth in his model, which makes it possible to model the grammar of the language by considering short-term memory. Nowadays, treebank resources make it possible to measure language performance on real data. Liu (2008) measured the dependency distance/length, which is the linear distance between a governor and a dependent, using dependency treebanks, and showed that the dependency distance has a tendency to be minimized (see also Tesnière, 1959: chapter 7; Futrell et al., 2015). Another typical question concerning the complexity of syntactic and cognitive ability concerns the limitation on the level of center-embedded constructions. Miller and Chomsky (1963) defined center-embedded constructions as a "nesting of dependencies, which occurs when X is embedded in another constituent Y, with material in Y to both the left and right of X," and remarked that increasing the levels of center-embedding makes the sentence incomprehensible. According to the psycholinguistic research by Lewis (1996), an English sentence has two levels of center-embedded clauses at most. In Japanese the total number can reach three (Lewis, 1996). This syntactic limitation is hypothesized to be related to the constraints of short-term memory. Kahane et al. (2017) considered the dependency flux, which is the set of dependencies linking a word on the left with a word on the right in a given position in the text and computed the flux weight, i.e. the maximum number of disjoint dependencies in the flux. As they showed, the flux weight, which measures the level of center-embedding constructions, is limited to 5 in the 70 treebanks of UD 2.0.

The dependency length carries only linear information and does not make it possible to measure the complexity of the configuration of dependencies, while the flux weight only evaluates the shape of the configuration of dependencies, without considering whether the dependencies in the configuration are long or short. We therefore propose a combined weight measure, in order to account for these two measurements at the same time. The calculation of dependency length, flux weight and combined weight will be presented in the next section.

2. Dependency flux

2.1. Flux size and flux weight

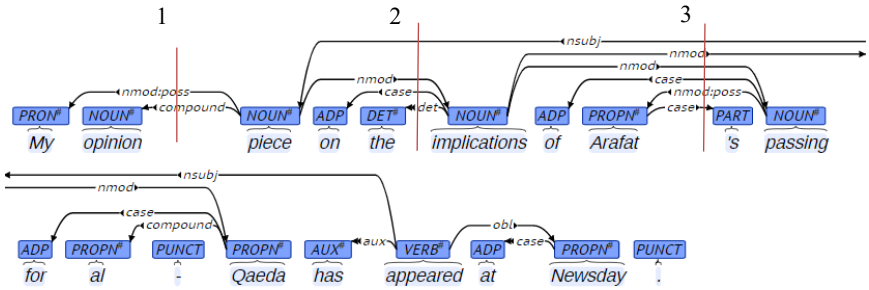

Figure 1. A dependency tree from UD-English-Original, with three positions considered

According to Kahane et al. (2017), dependency flux is the set of dependencies linking a word on the left with a word on the right in a given position. In Figure 1, three examples of flux positions are indicated by a vertical line: position 1 (opinion, piece), position 2 (the, implications), and position 3 (Arafat, 's). The flux size is the number of dependency links crossing the position. For position 1, we have two links, labeled *nmod:poss* and *compound*, that link a word to the left and a word to the right and the flux size is 2; for position 2, the flux size is 4; for position 3, the flux size is 6.

A set of dependencies is said to be disjoint if the dependencies do not share any vertex. The number of disjoint dependencies measures the center-embedding level (Kahane et al., 2017). For instance in position 2, there are two disjoint dependencies, [appeared -*nsubj*-> piece] and [implications -*case*-> on], which do not share any vertex, and represent exactly a center-embedded construction from the point of view of constituency analysis: [piece [on the implications] appeared]. In position 3, we find a set of four disjoint dependencies: [appeared -*nsubj*-> piece], [implications -*nmod*-> Qaeda], [passing -*case*-> of] and [Arafat -*case*-> 's]. In this position, we find a more

complicated center-embedded construction: [piece on the [implications [of [Arafat 's] passing] for Qaeda] appeared]. The flux weight is the size of the largest disjoint sub-flux. For position 1 the weight is 1, 2 for position 2, 4 for position 3.

2.2. Granularity

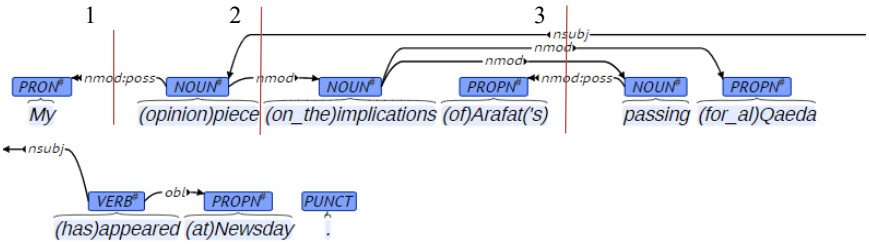

Figure 2. Aggregated tree, tokens in parenthesis are ignored.

At the modeling level, functional relations do not have a unified behavior in every treebank, and some of them are language specific relations. For example, the relation *clf* (classifier) exists in only a few languages, such as Chinese, and can form an additional disjoint dependency in comparison with other languages. It is possible to adjust the granularity of the syntactic analysis in order to make the different treebanks more comparable, for example by keeping only the content words and eliminating relations of the kind: auxiliary, case, conjunction, non-personal relations such as expletives, determiners, and parataxis.

This gives us an aggregated tree, such as the one in Figure 2. The three positions considered in Figure 1 are still marked by a vertical line. In position 1, there remains only one relation in the flux. In position 2, there are no longer any disjoint dependencies. The flux weight in position 1 and position 2 is now 1. For position 3, we have 3 disjoint dependencies, [appeared *-nsubj->* piece], [implications *-nmod->* Qaeda] and [passing *-nmod :poss->* Arafat] and the flux weight is then 3.

3. Dependency flux combined with dependency length

Our hypothesis of the complexity for sentence processing considers two aspects. On the one hand, the complexity depends on the number of disjoint dependencies that we measure by flux weight; on the other hand, it depends on the dependency length (modulo granularity). Thus, by combining the length of dependencies and the flux weight, we introduce a new measure, which we call the *combined weight* of the flux. We would like to look at how the combined

weights behave among the treebanks, as well as to study the characteristics of this new measure.

The combined weight in a given position is the sum of the dependency length of the longest disjoint dependencies. In the aggregated tree of Figure 2, for position 1, we have only one dependency, the length of which is 1, so the combined weight is W_c=1; for position 2, W_c=5; for position 3, W_c= 9 = 5 [appeared-*nsubj*->piece] + 3 [implications-*nmod*->Qaeda] + 1 [passing-*nmod:poss*->Arafat].

3. Results and discussion

3.1. Granularity

By calculating the flux weight of all inter-word positions for every treebank, we found that the maximum weight varies between 3 and 5 in the model of aggregated trees. It is 3 for Vietnamese and Slovak, 10 languages have a maximum weight of 5 and the other 25 treebanks have a flux weight of 4. In comparison with the original treebanks, where the maximum weight varies between 4 (4 treebanks) and 6 (12 treebanks), our model of aggregated trees brings the maximum weight of different treebanks closer. We also obtain the same result for the average weight.

3.2. Combined weight

As shown in Figure 3, the average weight of aggregated trees (AT) is stable and is more universal, because it only considers the center-embedding levels. The average combined weight of aggregated trees (AT) shows slightly the same trend, but it accentuates the differences among the treebanks. The combined weight carries more information about syntactic complexity.

The dependency length takes linear information into account, is correlated with sentence length, and is sensitive to genre (Jiang & Liu, 2015). As we lack information about genre, we cannot determine whether this difference in combined weight is due to different types of languages or to different genres.

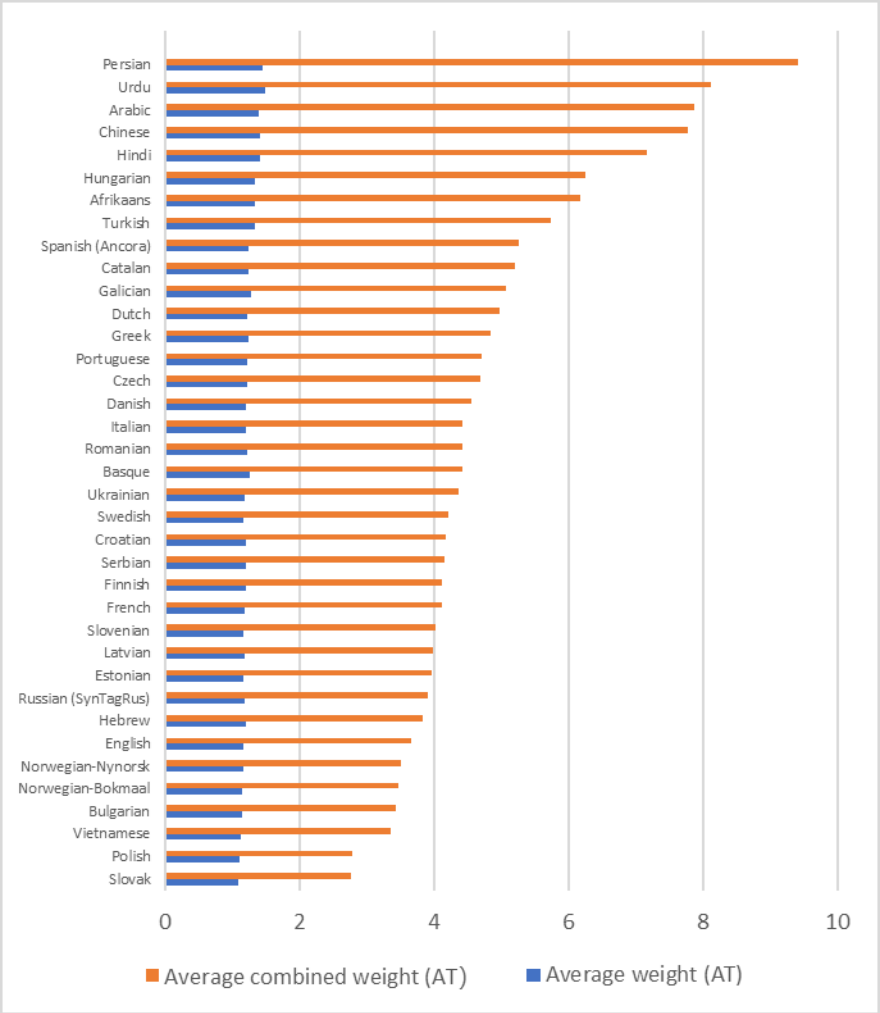

Figure 3. Average combined weight of aggregated trees (AT) and average weight of aggregated trees (AT) in 37 languages. (For more information see the supplementary materials.)

References

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33), 10336-10341.

Jiang, J., & Liu, H. (2015). The effects of sentence length on dependency distance, dependency direction and the implications–Based on a parallel English–Chinese dependency treebank. *Language Sciences*, *50*, 93-104.

Kahane, S., Yan, C., & Botalla, M. A. (2017). What are the limitations on the flux of syntactic dependencies? Evidence from UD treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)* (pp. 73-82).

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25*(1), 93-115.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159-191.

Miller, G. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. Psychological Review 63. 81–97.

Miller G. A, Chomsky, N. (1963). Finitary models of language users.

Tesnière, L. (1959). Eléments de la syntaxe structurale. Paris: Klincksieck.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, *104*(5), 444-466.