Interactive Workshop Measuring Language Complexity, Freiburg 2019

**Locus of marking and dependency length in possessive noun phrases**

Description of methods 30 June 2019

Kaius Sinnemäki and Viljami Haakana

General linguistics, University of Helsinki

[kaius.sinnemaki@helsinki.fi](mailto:kaius.sinnemaki@helsinki.fi)

1. Which level of language (e.g. morphology, syntax, phonology, morphosyntax) is addressed?

We approach both morphological and syntactic complexity in possessive noun phrases by using the Universal Dependencies (UD) data. First, we measure **morphological complexity** by approaching locus of marking in possessive noun phrases. Locus of marking can be interpreted also as a measure of morphosyntactic complexity, because it tracks morphological marking of syntactic relations. But since it first and foremost tracks different types of morphological marking, we treat our measure as representing morphological complexity. Second, we approach **syntactic complexity** by measuring dependency length in possessive noun phrases.

2. What exactly is measured (e.g. irregularity, transparency, lexical diversity)?

Morphology. Locus of marking refers to the position of morphological marking of syntactic relations in a construction (e.g., Nichols 1992; among many others). There are four logical loci for morphological marking: it occurs either on the head of the construction (the possessed in possessive NPs), the dependent of the construction (the possessor), on both (called double marking) or on neither (called zero marking). A fifth type, floating marking, is excluded here, since to our knowledge none of the sample languages use it in possessive noun phrases.

We delimit the analysis to constructions in which the possessed is a full noun, that is, either a common or proper noun (UPOS NOUN and PROPN, respectively). We delimit the analysis to constructions in which the possessor is personal possessive pronoun (e.g., *my house*), common noun (e.g., *the house of men*), or proper noun (e.g., *John’s house*); however, we exclude demonstrative and other pronouns as possessors.

We define dependent marking as any morphological marking on the dependent, be it via suffixes, tones, morphophonological alternations, clitics, or independent function words. The same definition applies to head marking.

Our measure focuses on the degree (or percentage) of head marking and the degree of dependent marking in possessive noun phrases. In other words, we calculate the proportion of possessive noun phrases in a language that use head marking and the proportion of possessive noun phrases in a language that use dependent marking. The rationale for this approach is to evaluate what is the relative importance of these two morphological types in marking syntactic relations in a language. It is thus very roughly related to the idea of functional load, which measures the relative importance of contrasts between linguistic units.

Syntax. Dependency length measures the distance between the syntactic head and its dependent in the construction. In possessive noun phrases this means the distance between the possessed (head) and the possessor (dependent). The distance is 1 if the head and dependent are adjacent to one another, but any intervening word between them increases distance by 1. In the CoNLL-U format content words, function words, and clitics are all treated as separate words that receive a unique word index (ID) in the sentence and thus potentially increase dependency length.

3. How the measure is operationalized/calculated?

Morphology. The process for operationalizing our approach has the following steps. First, we measure the presence vs. absence of the four different types of morphological marking in possessive noun phrases. On top of these four types we distinguish two subtypes of head marking, those in which the dependent is present in the construction as an independent syntactic constituent and those in which the dependent is absent as a separate syntactic constituent. Second, we then count the frequency of each type. These types are here labelled as Nsubtype , where subtype is one of the following: dep = dependent marking; head = head marking; double = double marking; zero = zero marking; head0 = head marking with no dependent as a separate syntactic constituent. Third, degree of dependent marking (dm in 6. Labeling) is then calculated as the proportion of those constructions in which there is any type of dependent marking as in (1):

Degree of head marking (hm in 6. Labeling) is calculated as the proportion of those constructions in which there is any type of head marking as in (2):

Syntax. Dependency length is first calculated as the absolute value of the subtraction of the head and dependent in each possessive noun phrase. We then calculate the average of the dependency lengths in each morphological type:

At this stage we have not yet analyzed all multiword possessors consistently, but only those languages for which we have written a separate script. In some languages we have been able to include those constituents whose syntactic function is flat, but in most languages we have not yet done this because it has proved to be very laborious. Any feedback in this regard is highly appreciated.

4. What the theoretical motivation for this measure is?

Morphology. Since the pioneering work of Johanna Nichols in the 1980s and 1990s the locus of marking has become an important approach in language typology for classifying languages in terms of their morphological properties. Earlier typological research that has been based on grammar descriptions has suggested that head and dependent marking have an inverse relationship (e.g., the work of Anna Siewierska, Dik Bakker, and Kaius Sinnemäki): languages tend to have either head marking or dependent marking but it is much less common for languages to have both of them or just zero marking. Our measures provide quantitative corpus data on how frequently these types are in actual language use; this approach enables assessing their relationship more concretely than in grammar-based typology.

Syntax. Dependency length has also become a highly important and useful measure of syntactic complexity. It provides an approach to measuring to what extent languages minimize distances between constituents that have a dependency relationship. Our motivation for measuring dependency length is to determine whether there are any differences in dependency length between the different types of morphological marking in possessive noun phrases. To our knowledge this has not been done before, at least for possessive noun phrases. If dependency length varies systematically according to the type of morphological marking, this would be interesting evidence for efficiency in the relationship between morphology and syntax.

5. What the advantages and the drawbacks of this measure are?

Morphology. The advantage of focusing on locus of marking in possessive noun phrases is that it provides a well-delimited functional domain instead of trying to evaluate language’s morphological complexity more comprehensively. Our approach measures the importance of each type of morphological marking in marking syntactic relations between the head and dependent in possessive noun phrases and thus it can be related to functional load. Our narrow focus to possessive noun phrases also provides possibilities of contrasting the results with grammar-based typological surveys and evaluating whether there is converging or contradicting evidence e.g. for complexity trade-offs in this area. Our hypothesis is that the inverse relationship between head and dependent marking found earlier in grammar-based surveys will become even clearer with quantitative corpus data. We thus expect that the evidence from corpus data will provide converging evidence for such trade-off.

On the other hand, measuring the proportion of certain type of morphological marking used in marking possessive noun phrases is not very directly a measure of complexity. But there does not seem to be an easy way around this issue. Complexity measures that have been used in corpus analysis earlier, such as entropy and type/token ratio, are not very informative when comparing just two types of morphological marking in a narrow grammatical area.

Syntax. In our approach we measure the length of dependency between possessor and possessed in constructions that use different type of morphological marking. One advantage of this approach is that it enables us to compare dependency length under different morphological types and thus evaluate whether dependency length correlates with presence vs. absence of a certain type of morphological marking. We expect to see shorter dependency length in possessive noun phrases with zero marking compared to those with some overt morphological marking.

We are not yet sure of the possible drawbacks of measuring dependency length in the way we propose. One potential problem is our definition of morphological marking: if independent function words are used for morphological marking and those words occur between the head and the dependent, they also increase dependency length. Currently we think this about right: they should increase dependency length. On the other hand, it might be useful also to contrast our current measure with one that excludes the function words for head and dependent marking from dependency length to determine whether it is just the function words that are responsible for possible differences.

6. Labeling of the measures

The following are the labels used in the results file Sinnemäki.csv:

* id: language/treebank id, provided by the workshop organizers
* language: language name, provided by the workshop organizers
* dm = percentage of possessive noun phrases that have dependent marking
* hm = percentage of possessive noun phrases that have head marking
* dep\_dl = dependency length in possessive NPs in which the relation between head and dependent is marked on the dependent
* double\_dl = dependency length in possessive NPs in which the relation between head and dependent is marked on both the head and the dependent
* head\_dl = dependency length in possessive NPs in which the relation between head and dependent is marked on the head only (excluding constructions in which there is no dependent present)
* zero\_dl = dependency length in possessive NPs in which the relation between head and dependent is marked on neither the head nor the dependent

References (to be completed)

Nichols, Johanna 1992. *Linguistic Diversity in Space and Time*. Chicago: The University of Chicago Press.