# A Quantitative and Typological Approach to Correlating Linguistic Complexity

Yoon Mi Oh  François Pellegrino  Egidio Marsico  Christophe Coupé
Laboratoire Dynamique du Langage, Université de Lyon and CNRS, France
{yoon-mi.oh, francois.pellegrino}@univ-lyon2.fr,
{egidio.marsico, christophe.coupe}@cnrs.fr

## Hypothesis and objectives

The equal complexity hypothesis states that "all human languages are equally complex" (Bane, 2008). Menzerath's law is well-known for explaining the phenomenon of self-regulation in phonology: "the more sounds in a syllable the smaller their relative length" (Altmann, 1980). Altmann, who made the mathematical formula of this law (Forns and Ferrer-i-Cancho, 2009), assumed that it can be applied to morphology as well - "the longer the word the shorter its morphemes" (Altmann, 1980) - and proved that the clause length depends on sentence length (Teupenhayn and Altmann, 1984).

Some previous works on morphological complexity (Bane, 2008; Juola, 1998) asserted that morphology is a good starting point for complexity computation for its clearness, compared to other more ambiguous domains such as semantics. The best-known method of calculating morphological complexity is to take the numbers of linguistic constituents into account (Bane, 2008; Moscoso del Prado, 2011), with different mathematical formula to be applied to these figures. The following two paradigms are commonly employed: i) information theory (Fenk et al., 2006; Moscoso del Prado et al., 2004; Pellegrino et al., 2011) ii) Kolmogorov complexity (Bane, 2008; Juola, 1998).

The main goal of our work is to explore interactions between phonological and morphological modules by means of crossing parameters of these two linguistic levels. This paper provides preliminary results obtained from a corpus-based cross-language study.

## Methodology and preliminary results

Our 14-language corpus is based on the Multext multilingual corpus (Campione and Véronis, 1998). For each language, 15 short texts which consist of 3-5 sentences translated from British English are recorded by 5 male and 5 female native speakers. The data of 6 languages (English (eng), German (deu), Italian (ita), Mandarin Chinese (cmn), Spanish (spa) and Vietnamese (vie)) are taken from the Multext corpus, and the data of the other 8 languages (Basque (eus), Catalan (cat), French (fra), Hungarian (hun), Japanese (jpn), Korean (kor), Turkish (tur) and Wolof (wol)) have been collected by the authors.

Two types of parameters are taken into account in this study. First, at the phonological level, and following Pellegrino et al. (2011), a set of phonological factors is employed. For each language, the *syllabic rate*, (the number of syllables pronounced per second), is computed. Additionally, using Vietnamese as an external reference, a *syllabic information density* (resp. *word information rate*) is defined for each target language as the average ratio between the total number of syllables (resp. words) in a text in Vietnamese and the number of syllables (resp. words) of this text translated in the target language.

Our method of measuring the information density computes the average amount of information carried by syllables and words at the text level. Thus, it differs from studies related to the principle of uniform information density (Frank and Jaeger, 2008), since the latter focus on the variation of information transmitted during communication. Figure 1 illustrates the negative correlation ($R^2$= 0.65) between these phonological factors, i.e. a trade-off between syllabic rate and information density (Pellegrino et al., 2011).

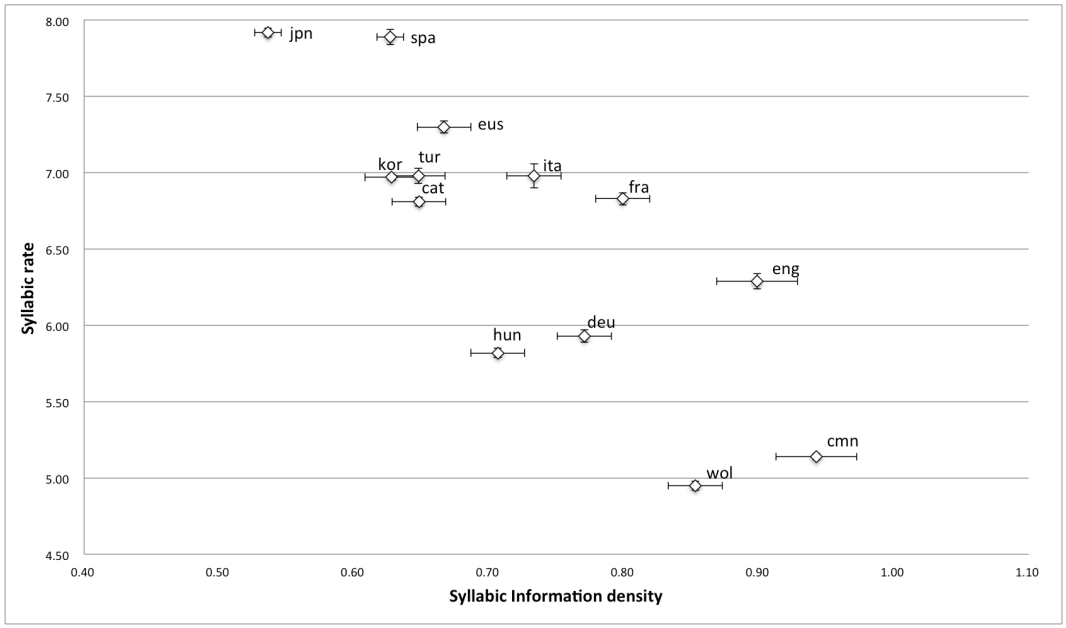


**Figure 1: Syllabic rate and syllabic information density (Error bars indicate standard error)**

Second, at the morphological level, the languages of our corpus can be classified into three categories, as shown in Table 1 (Greenberg, 1960).

| Category | Languages |
|---|---|
| **Agglutinative languages** | Basque, Hungarian, Japanese, Korean, Turkish |
| **Fusional languages** | Catalan, English, French, German, Italian, Spanish, Wolof |
| **Isolating languages** | Mandarin Chinese, Vietnamese |

**Table 1: Morphological classification**

In order to investigate the relations between phonological and morphological modules, we compare the average number of syllables per word and the information density calculated at the word level and at the syllable level, respectively in Figures 2 and 3.

Figure 2 exhibits a strong positive correlation ($R^2$= 0.84) between the average number of syllables per word and the information density at the word level, which logically means that the longer the word, the more information it contains. In general, there are more syllables per word in agglutinative languages (in black) than in fusional languages (in grey). Chinese as an isolating language is marked in white. Furthermore, Figure 3 shows that at the syllable level, fusional languages have a tendency towards higher information density compared to agglutinative languages.

Values of languages in the same morphological category are quite dispersed. In Figure 2, regarding fusional languages, a large difference exists, for example, between German with a very complex declension system and English with a limited morphological system (Moscoso del Prado, 2011). Japanese, which has a relatively simple phonological system, has the largest number of syllables per word and transmits the least amount of information per syllable (Figure 3). Compared to Japanese, Mandarin Chinese, an isolating language with a relatively complex phonological system, shows completely opposite values.

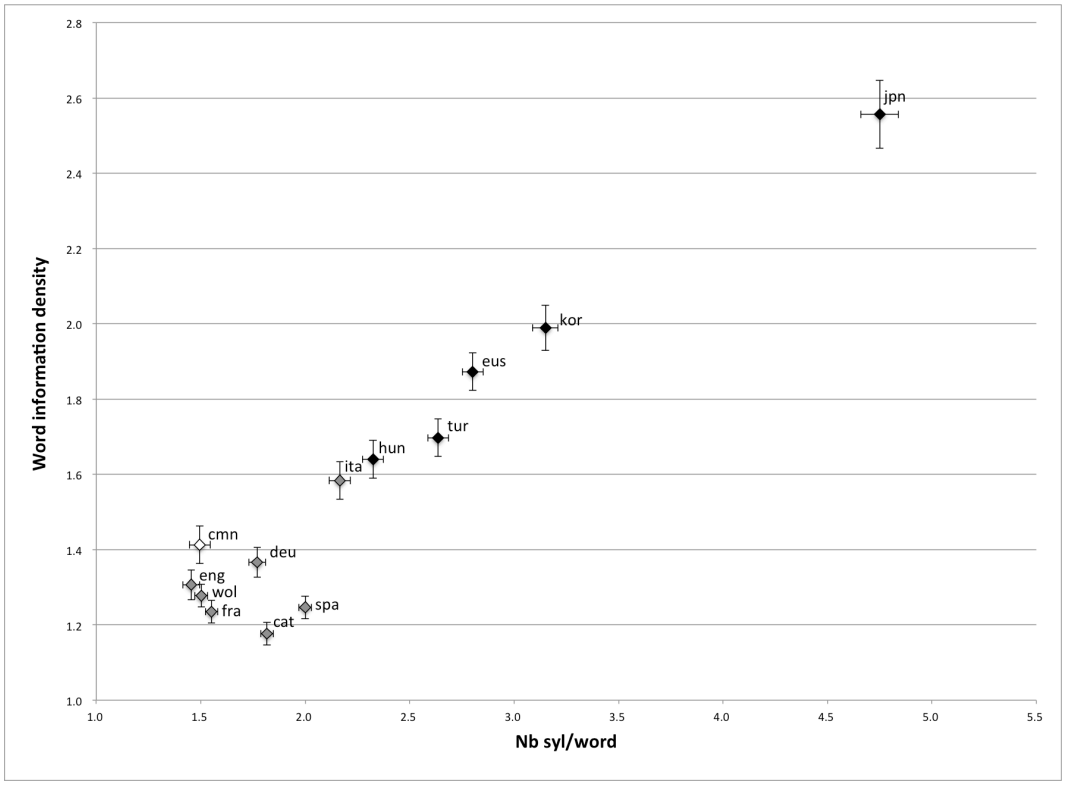


**Figure 2: Word information density and mean number of syllables per word (Error bars indicate standard error)**

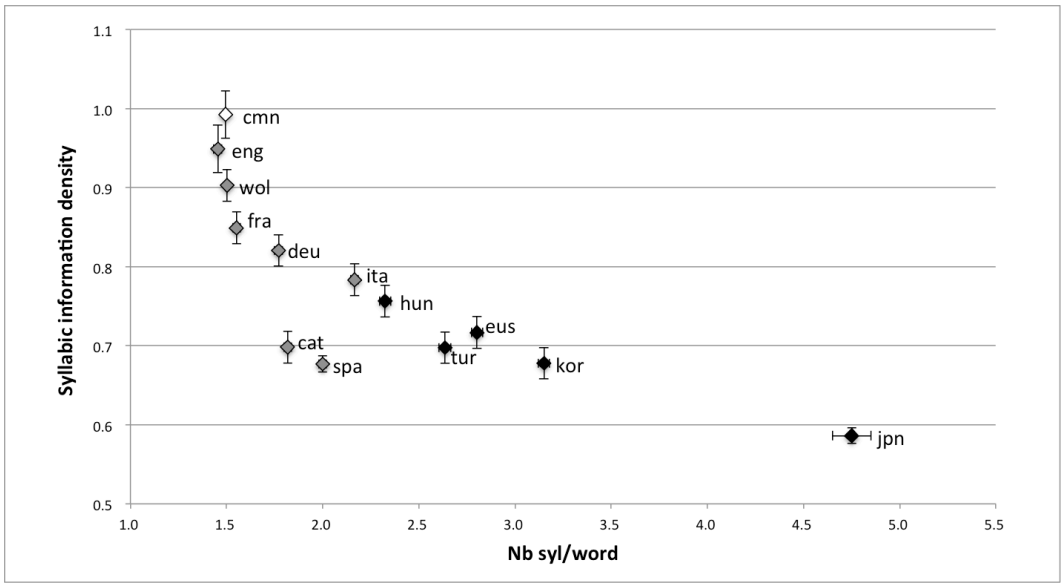


**Figure 3: Syllabic information density and mean number of syllables per word (Error bars indicate standard error)**

## Discussion and further work

Fenk et al. (2006) defined word complexity as the mean number of syllables per word and syllable complexity as the mean number of phonemes per syllable, and found a negative linear correlation between these two figures. Similarly, our result shows a negative corre-

lation between word complexity and information density at the syllable level, i.e. the less complex a word, the more information per syllable.

Furthermore, according to our results, despite the dispersed values of languages in the same morphological category, some differences are observed between these categories and agglutinative languages clearly tend to have longer words than fusional languages. Fenk-Oczlon and Fenk (1985) showed that the average number of syllables per clause depends on the mean number of phonemes per syllable, but the analysis at word level had not been done before.

These preliminary results show a relation between the morphological and phonological modules. In further studies, this relation will be investigated in more details by analyzing our multilingual parallel data, and by adding more isolating languages to observe their pattern. We are currently working on unsupervised morpheme segmentation, using Morfessor (Creutz and Lagus, 2005), in order to compare our multilingual data at morpheme level. At the same time, we aim to compare the average number of words per sentence in order to correlate the linguistic complexities of three different levels.

## References


Altmann, G. 1980. Prolegomena to Menzerath's law. *Glottometrika*, 2, 1–10.

Bane, M. 2008. Quantifying and measuring morphological complexity. In *Proc. of the 26th West Coast Conference on Formal Linguistics*, 69-76.

Campione, E. and Véronis, J. 1998. A multilingual prosodic database. In *Proc. of ICSLP98*, Sydney, Australia, 3163-3166.

Creutz, M., and Lagus, K. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0*. Publications in Computer and Information Science, Report A81, Helsinki University of Technology, March. URL: http://www.cis.hut.fi/projects/morpho/

Fenk, A., Fenk-Oczlon, G. and Fenk, L. 2006. Syllable complexity as a function of word complexity. *In The VIII-th International Conference "Cognitive Modeling in Linguistics" Vol. 1*, 324-333.

Fenk-Oczlon, G., and Fenk, A. 1985. The mean length of propositions is 7 plus minus 2 syllables–but the position of languages within this range is not accidental. In *Proc. of the XXIII International Congress of Psychology: Selected/Revised Papers*, Vol. 2, 355–359.

Forns, N. and Ferrer-i-Cancho, R. 2009. The self-organization of genomes. *Complexity, 15(5)*, 34-36.

Frank, A., and Jaeger, T. F. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proc. of the Cognitive Science Society*.

Greenberg, J. 1960. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics, 26(3)*, 178-194.

Juola, P. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, *5*(3), 206–213.

Moscoso del Prado, F., Kostić, A., and Baayen, R.H. 2004. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition, 94(1),* 1-18.

Moscoso del Prado, F. 2011. The Mirage of morphological complexity, In *Proc. of the 33rd Annual Conference of the Cognitive Science Society*, 3524-3529.

Pellegrino, F., Coupé, C., and Marsico, E. 2011. A cross-language perspective on speech information rate. *Language, 87(3),* 539–558.

Teupenhayn, R., and Altmann, G. 1984. Clause length and Menzerath's law. *Glottometrika 6,* 127-138.