Linguistic Complexity and Information Rate: Quantitative Approaches

Yoon Mi OH

Laboratoire Dynamique du Langage - UMR 5596, Université de Lyon 2, yoon-mi.oh@univ-lyon2.fr

Objective and Hypothesis

What?

Main goal of the present project is to investigate the **relations** between **linguistic complexity** and **information rate.**

Why?

According to the **cross-linguistic** research in the laboratory **DDL** (Pellegrino et al., 2011), there is a **negative correlation** between **syllable complexity** and **speech rate**. ➡ The **more complex** the syllable, the **slower** the transmission of information.

How?

➡ by adding **more languages** with various **syllable structure** and **phonological inventory.**

➡ by analyzing **multilingual oral** and **text corpus** of 12 languages.

About the corpus

For analyzing each language, **two types of corpus** are required.

① Oral corpus

▹ made up of **20 texts translated** from the original texts in English with slight modification if necessary. ➡ **The same semantic information**

▹ 10 native speakers (5M & 5F) are recorded for each language.

② Text corpus

▹ large amount of **plain text corpus** which contains more than **60k words**, in order to get **a usage-based syllable frequency list**.

Methodology

① Oral corpus

▹ for calculating **syllable rate** (**SR**: number of syllables uttered per second), **information density** (**ID**: amount of linguistic information per syllable) and **information rate** (**IR**: amount of information transmitted per unit of time).

❗ **Silence intervals** longer than 150ms were removed.

❗ In case of **information density** and **information rate**, corresponding values were calculated respectively by **pairwise comparisons** of the length of data (**number of syllables**) and the **mean duration** of data, using **Vietnamese** as an external reference.

② Text corpus

▹ for calculating **syllabic inventory** (**SI**), **syllable complexity** (**SC**) and **syllabic entropy** (**H**: cognitive cost of using a syllable (Ferrer i Cancho et al., 2007)).

▹ Automatic **syllabification** by specific rules for each language ➡ syllable frequency list ➡ syllable inventory and syllable complexity

▸ **Syllable complexity** (**SC**): number of **syllable constituents**

▸ **Syllabic entropy** (H_L) ➡ $H_L = -\sum_{i=1}^{N_L} p_i \log_2(p_i)$

(N_L = syllable inventory, i = each syllable, p_i = frequency of each syllable)

Preliminary results

① Comparison of information density, syllable rate and information rate

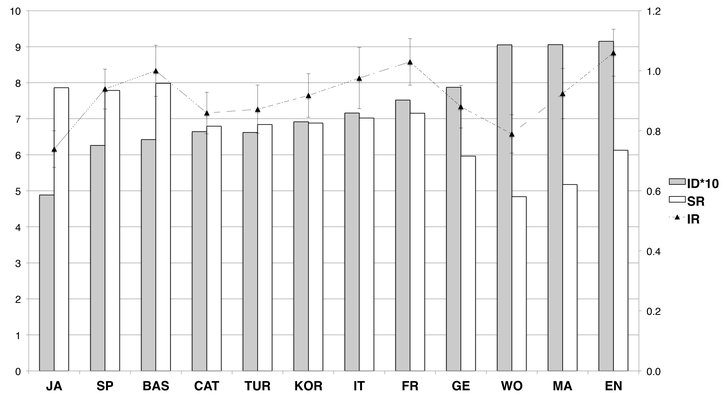

Figure1: Comparing information density (**ID**), syllable rate (**SR**) & information rate (**IR**) of 12 languages (JA: Japanese, SP: Spanish, BAS: Basque, CAT: Catalan, TUR: Turkish, KOR: Korean, IT: Italian, FR: French, GE: German, WO: Wolof, MA: Mandarin, EN: English)

Figure 1 shows a **comparison** of information density, syllable rate (left axis for both) and information rate (right axis) and illustrates **similar information rate** values regardless of distinct differences between their information density and syllable rate values.➡ **Trade-off** between information **density** and syllable **rate**

② Relation between information density and syllabic entropy

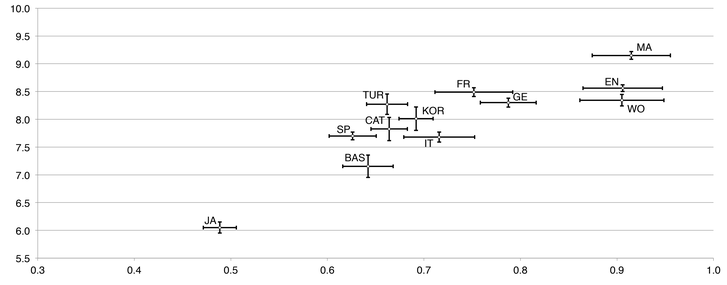

Figure2: Correlation between **information density** and **syllabic entropy**

The **strongest correlation** is observed between **information density** (**ID**, x-axis) and **syllabic entropy** (**H**, y-axis) (Pearson's cor = 0.86, p-value = 0.0004, Spearman's rho = 0.81, p-value = 0.002).

➡ It reveals that there is a **close correlation** between **syntagmatic dimension** (information density: the encoding of linguistic information) and **paradigmatic dimension** (syllabic entropy: the distribution of syllable frequencies) of **linguistic complexity**.

Conclusions and further work

By adding **more languages** with distinctive phonological features to our project, we aim to observe a negative **correlation** (trade-off) between **information density** and **syllable rate**, which **regulates information rate** in our hypothesis.

In future, the notion of **complexity** which is currently limited to **phonological** level will be expanded to **morphosyntactic** level.

References

Ferrer i Cancho, R., & Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2007, P06009.

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

❗ You can also download a **PDF-version** of this poster on my personal page of DDL website at **www.ddl.ish-lyon.cnrs.fr** or **here**. ➡