Quantitative Approaches to Phonological Complexity: the Case of East Asian Languages

Yoon Mi OH & François PELLEGRINO
Dynamique du Langage Laboratory

KACL 2012,
December 10, 2012

Overview

1. Theoretical framework

2. Corpus description and analysis method

3. Results

4. Perspectives

Overview

1. Theoretical framework

-> Human language is a “complex system”!

2. Corpus description and analysis method

3. Results

4. Perspectives

Our hypothesis

- Human language is a **complex system**.
  : trade-off, balance, self-organization.
- **Information theory** point of view

: A trade-off (self-organization) exists between the speech rate and the information density in human communication, regardless of their coding system (Pellegrino et al. 2011).

Phonological complexity

- Two ways of measuring phonological complexity

- **Linguistic approach**

: Average syllabic complexity in terms of number of constituents (number of segments + tone).

- **Quantitative approach**

: “Syllabic entropy” (calculated from the distribution of syllable frequencies), notion adapted from the Information theory.

Overview

1. Theoretical framework

2. Corpus description and analysis method

-> Multilingual oral and text corpus of East Asian languages

3. Results

4. Perspectives

Multilingual Oral corpus

- **Description**

- Subset of multilingual oral corpus in Japanese, Korean, Mandarin supplied by EUROM 1 corpus extracted for the MULTEXT project (Campione & Véronis (1998), Komatsu et al. (2004), Kim et al. (2008)).

- 20 short texts (of 3-5 semantically connected sentences) translated in each language with local adaptation when necessary.

- 6 speakers for Japanese , 10 for Korean, 9 for Mandarin.

Example of oral script

Japanese

Passage: 01

1. 家の浄水器の調子が悪いです。

2. 水圧が高すぎるみたいで、排水口からずっと水滴がたれています。

3. すみませんが、火曜日の午後に技術者派遣の手配をしていただけますか？

4. 今週は火曜日しか都合がつけられないのです。

5. 念のために書面にて手配確認してもらえるとありがたいです。

Nb of syllables: 120

Korean

Passage: 01

1. 연수기가 고장이 났습니다.

2. 수위가 너무 높아서 물이 계속 넘치거든요.

3. 다음주 화요일 아침에 사람을 좀 보내주실 수 있으세요?

4. 제가 다음주는 그날 밖에 시간이 안되거든요.

5. 정확한 일정을 메일로 보내주시면 감사하겠습니다.

Nb of syllables: 89

Mandarin

Passage: 01

1. 我的净水器出毛病了。

2. 水位太高，所以水总是流出来。

3. 您能不能派人星期二早上来看一下？

4. 这星期我只有那天有空。

5. 来之前最好能先来个电话。

Nb of syllables: 57

- **Basic notions**

- **Syllable rate**
  : Number of syllables uttered per second.
- **Information density**
  : Amount of linguistic information per syllable.
- **Information rate**
  : Amount of information transmitted per unit of time.

- **Analysis method**

- **Syllable rate** is calculated by removing silence intervals longer than 150ms.
- **Information density** and **information rate** are calculated respectively by pairwise comparisons of the total number of syllables per each text and the mean duration of data, using Korean as a reference.

Multilingual text corpus

- **Description**

- Large text corpus (internet, newspapers, books, etc) which are available online.

- Different resources for each language.
  - Japanese: Tamaoka and Makioka, 2004.
    (# of different syllables: 416, total # of syllables: 575.7M)
  - Korean: Kang Seung-Shik, Kookmin nlp corpus.
    (# of different syllables: 2026, total # of syllables: 31.2M)
  - Mandarin: PhD, Peng Gang, 2005.
    (# of different syllables: 1191, total # of syllables: 138M)

- **Analysis**

- Information theory-based approach

  : Language *L* is a source of linguistic sequences composed of syllables (σ) from a finite set (N_L) (Pellegrino 2012).

- Syllabic entropy: $$H_L = -\sum_{i=1}^{N_L} p_{\sigma_i} \log_2 \left(p_{\sigma_i} \right)$$

  - Cognitive cost of using a syllable (Ferrer i Cancho & Díaz- Guilera 2007)
  - Quantity of information of a syllable
  - probability ↓ information (entropy) ↑
  - probability ↑ information (entropy) ↓
  - p=1, no information

Overview

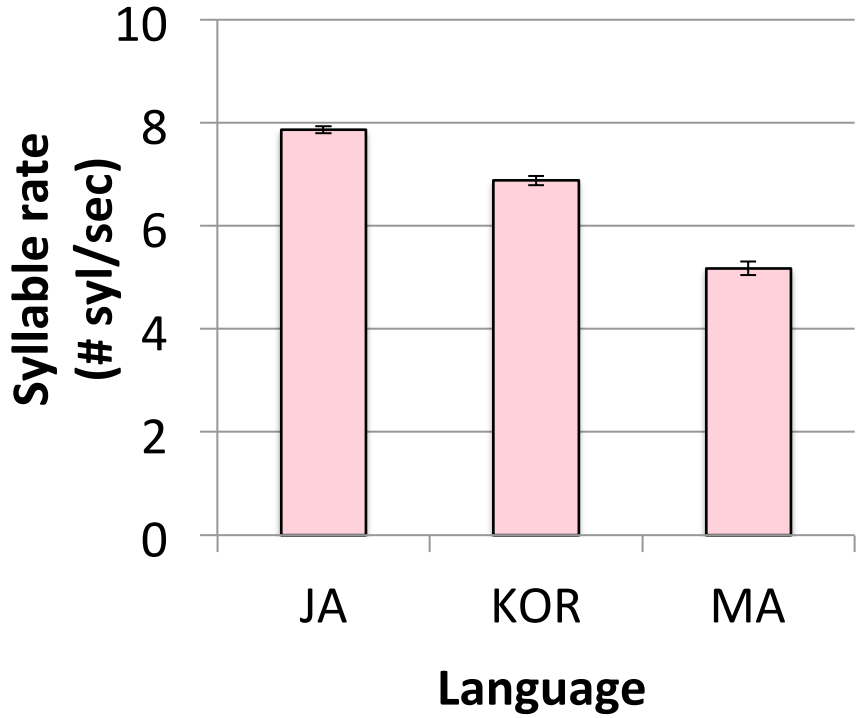
1. Theoretical framework

2. Corpus description and analysis method

3. Results

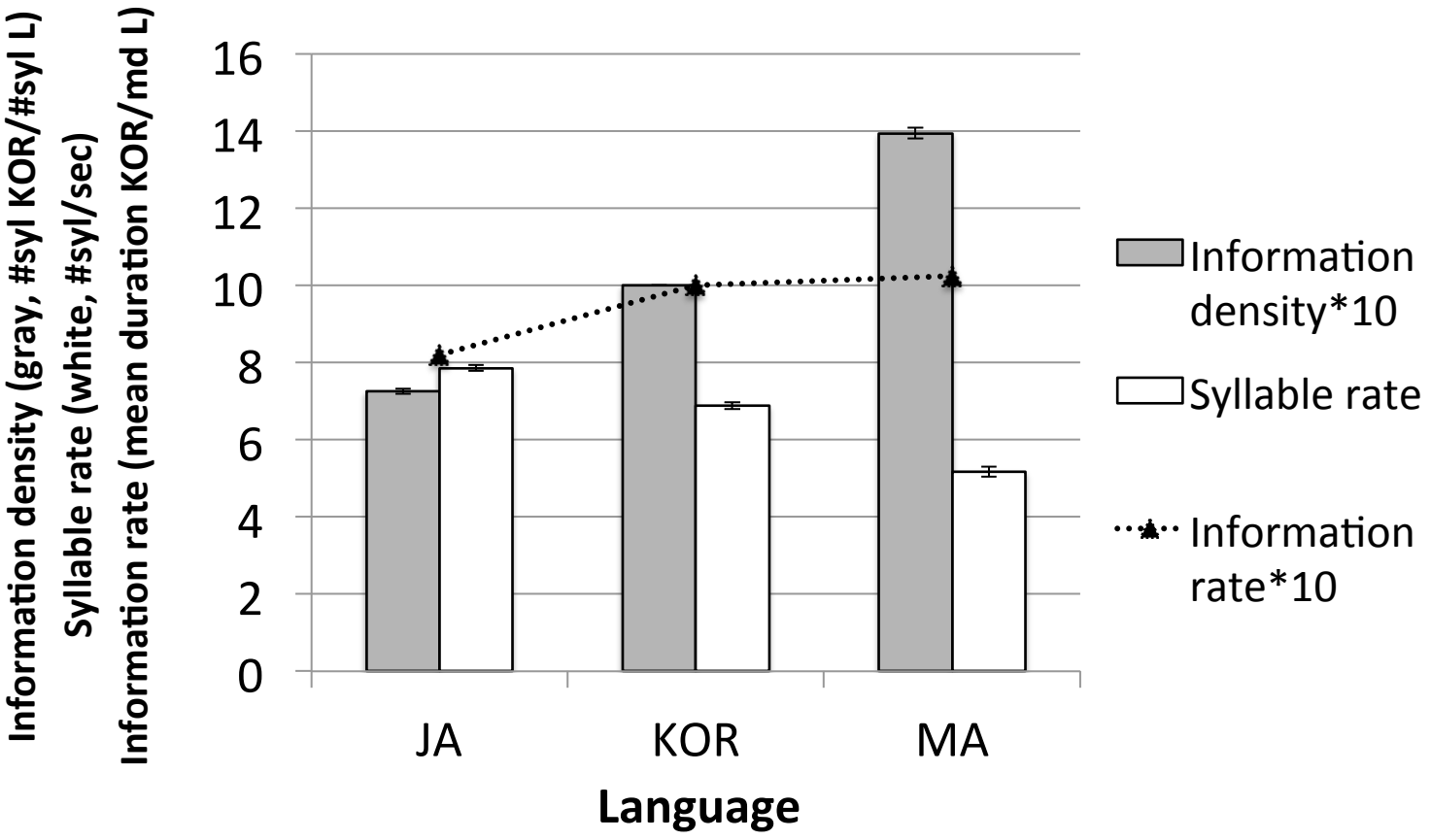
4. Perspectives

- Syllable rate of Japanese, Korean & Mandarin



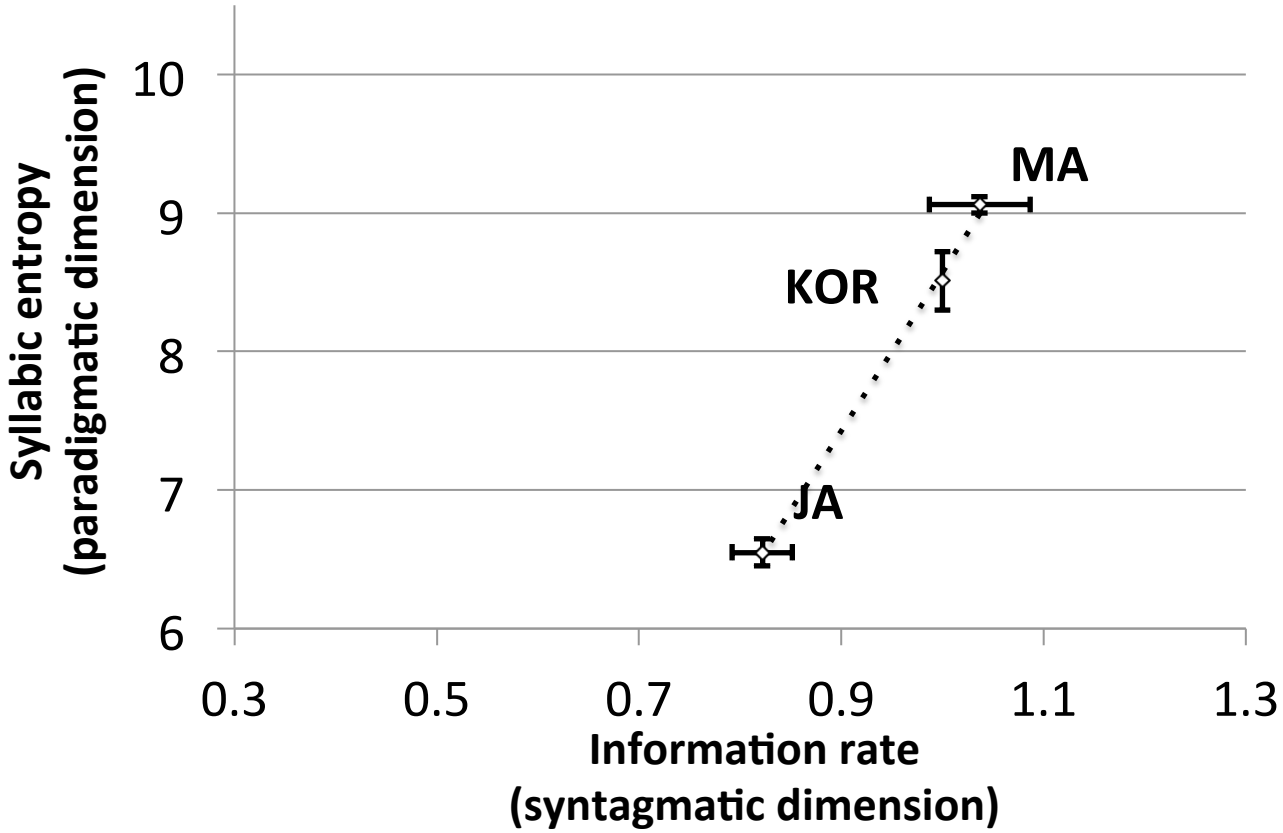
Language	Syllable rate	Confidence interval
JA	7.86	0.07
KOR	6.88	0.09
MA	5.18	0.13

- Information density, syllable rate & information rate of Japanese, Korean & Mandarin



-> **Negative correlation (trade-off)** between information density and syllable rate, regardless of information rate which varies little.

- Relation between information rate and syllabic entropy



-> Syntagmatic dimension (information rate) and paradigmatic dimension (syllabic entropy) of phonological complexity are related.

- **In conclusion**

- Our hypothesis: trade-off between syllable rate and information density -> stable value of information rate.

- Syllabic entropy: efficient method for computing phonological complexity -> no need to count the # of syllable constituents.

- Adding “1” for the tone in case of Mandarin, without taking the pitch accent into account in case of Japanese.

- Syllabic entropy/phonological complexity (paradigmatic dimension) and information rate (syntagmatic dimension) can be positively correlated -> need to add more languages to verify it!

Overview

1. Theoretical framework

2. Corpus description and analysis method

3. Results

4. Perspectives

Perspectives

- Language universal?

  : To prove our hypothesis (trade off between syllable rate and information density, which regulates information rate) -> add more typologically distant languages (14 languages for now: Bas, Cat, En, Fa, Fr, Ge, Hu, It, Ja, Kor, Ma, Sp, Tur, Wo).

- Study of syllable rates of bilinguals (Basque-Spanish and Catalan-Spanish speakers in Spain)

- Expansion of the notion of complexity to morphological and syntactic level.

References

Campione, E. & Véronis, J. (1998). A multilingual prosodic database. *Paper presented at the 5th International Conference on Spoken Language Processing*, Sydney: Australia.

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*, P06009.

Kim, S. Hirst, D., Cho, H., Lee, H., & Chung, M. (2008). Korean MULTEXT: A Korean Prosody Corpus.

Komatsu, M., Arai, T., & Sugarawa, T. (2004). Perceptual discrimination of prosodic types. *Paper presented at the Speech Prosody*, Nara: Japan.

Pellegrino, F., Coupé, C. & Marsico, E. (2011). A cross-language perspective on speech information rate, *Langage,* 87:3.

Pellegrino, F. (2012). Syllabic information rate: a cross-language approach, Dartmouth College, September, 27 2012.

감사합니다!

Merci beaucoup!