Grammar Efficiency and the One-Meaning–One-Form Principle

Relja Vulanović
Department of Mathematical Sciences
Kent State University at Stark
North Canton, Ohio, USA

QUALICO 2018

Overview

- A measure of how much Anttila's (1972) One-Meaning – One-Form Principle (**the Principle**) is violated has been proposed in (Vulanović & Ruff, QUALICO 2016).
- The measure is now incorporated in a new formula for calculating grammar efficiency.
- This is exemplified by parts-of-speech (PoS) systems in the sense of (Hengeveld, 1992).

Contents

- Measures of the degree of violation of the Principle
- Hengeveld’s PoS systems
- The old grammar-efficiency formula
- The new grammar-efficiency formula
- Results
- Conclusions

Notation

- $|A|$ is the number of elements of a finite non-empty set A.
- X = set of meanings, Y = set of forms
- Set of pairs (relation): $\Phi \subseteq X \times Y$
- B = set of one-to-one pairs:

$$B = \{(x, y) \in \Phi : \ \xi(y) = \upsilon(x) = 1\}$$

$$\xi(y) = |\{x \in X : (x, y) \in \Phi\}|, y \in Y$$
$$\upsilon(x) = |\{y \in Y : (x, y) \in \Phi\}|, x \in X$$

Basic Facts

- $|B| \leq |\Phi|$
- If Φ is a bijection (a one-to-one correspondence) between *X* and *Y*, then $|X| = |Y| = |\Phi| = |B|$.

The Measure $\mu(\Phi)$ 1

1. $\mu(\Phi) = 1$ if Φ is a bijection; otherwise $\mu(\Phi) > 1$
2. $\mu(\Phi)$ is greater if $|\Phi|$ is greater and if $|X|$ and $|Y|$ are smaller
3. $\mu(\Phi)$ is smaller if $|B|$ is greater
4. $\mu(\Phi) = \mu(\Phi^{-1})$, $\Phi^{-1} = \{(y, x) : (x, y) \in \Phi\}$

The Measure $\mu(\Phi)$ 2

- QUALICO 2016

$$\mu(\Phi) = \mu_\theta(\Phi) \coloneqq \frac{(1+\theta)|\Phi| - \theta|B|}{\min\{|X|, |Y|\}}, \theta > 0$$

- A simplified formula considered here:

$$\mu(\Phi) = \frac{|\Phi| - |B|}{\min\{|X|, |Y|\}} + 1 \geq 1$$

- Properties 1-4 satisfied.

The Weighted Formula

$$\mu(\Phi) = \frac{\|\Phi\| - \|B\|}{\min\{\|X\|, \|Y\|\}} + 1$$

- $\|A\| = w_1 + w_2 + \cdots + w_n,\ \ |A| = n$
  $$w_i = w_i(A), \qquad \min w_i = 1$$
- If $w_1 = w_2 = \cdots = w_n = 1$, then $\|A\| = |A|$ and v.v.

PoS Systems: Propositional Functions

- X = set of propositional functions (syntactic slots):
  P = head of predicate phrase
  R = head of referential phrase
  r = optional modifier of referential phrase
  p = optional modifier of predicate phrase
- $|X| = l =$ number of propositional functions in a PoS system, $1 \leq l \leq 4$.

Y = set of word classes, $|Y| = k$

Word class	P	R	r	p
Verbs	V	-	-	-
Nouns	-	N	-	-
Adjectives	-	-	a	-
Manner adverbs	-	-	-	m
Heads	H	H	-	-
Predicatives	₱	-	-	₱
Nominals	-	₦	₦	-
Modifiers	-	-	M	M
*	X_1	-	X_1	-
*	-	X_2	-	X_2
Non-verbs	-	Λ	Λ	Λ
*Non-nouns	Z	-	Z	Z
*	X_3	X_3	X_3	-
*	X_4	X_4	-	X_4
Contentives	C	C	C	C

Weights

- Weight of $\mathrm{P} = \alpha$
- Weight of $\mathrm{R} = \beta$
- Weight of $\mathrm{r} = \gamma$
- Weight of $\mathrm{p} = \delta$

l	Propositional functions in the PoS system
4	P R r p
3	P R r
3	P R p
2	P R
1	P

$$\alpha = 2.5, \beta = 2, \gamma = \delta = 1$$

Weights

- $\|X\|$ is the sum of weights of propositional functions:

  $\|X\| = \alpha + \beta + l - 2$ if $l = 2,3,4$;
  $\|X\| = \alpha$ if $l = 1$

- Weights of Φ (same for B):
  If $(x, y) \in \Phi$, its weight is defined as $w(x)w(y)$.

Weights of Word Classes 1

- For $y \in Y$, define $w(y)$ as the number of *horizontally and vertically connected* cells in the scheme

	Head	**Modifier**
Predication	-	-
Reference	-	-

Weights of Word Classes 2

- For instance, $w(\Lambda) = 3$

	Head	Modifier
Predication	-	Λ
Reference	Λ	Λ

- $w(\mathrm{X}_1) = 3$

	Head	Modifier
Predication	X_1	-
Reference	-	X_1

- Flexibility of word classes is penalized.

PoS System Types

- **Rigid** PoS systems ($k = l, \mu = 1$): VNam, VNaØ, VNØm, VNØØ, VØØØ (word classes are listed in the order which corresponds to the PRrp order of propositional functions they convey)

- **Flexible** PoS systems: $k < l$

Flexible PoS System Types, $l = 2,3$

l	k	PoS system type	μ
2	1	HHØØ	5.500
3	2	V₦₦Ø	3.000
		₱NØ₱	3.333
		VX_2ØX_2	3.250
		X_1NX_1Ø	3.625
		HHaØ/HHØm	4.000
	1	$X_3X_3X_3$Ø/X_4X_4ØX_4	6.500

Flexible PoS System Types, $l = 4$

l	k	PoS system type	μ
4	3	VNMM	2.000
		V₦₦m	2.500
		₱Na₱	2.750
		VX_2aX_2	2.800
		X_1NX_1m	3.100
		HHam	3.250
	2	VΛΛΛ	4.000
		ZNZZ	4.375
		₱₦₦₱/HHMM	4.250
		$X_4X_4aX_4/X_3X_3X_3m$	5.125
		$X_1X_2X_1X_2$	4.250
	1	CCCC	7.500

Absolute Grammar Efficiency

$$AE = Q\,\frac{|\text{Information}|}{|\text{Conveyors}|} = Q\,\frac{|X|}{|Y|} = Q\,\frac{l}{k}$$

The coefficient of proportionality *Q* depends on the complexity of the grammatical rules transforming the input *Y* to the output *X*:

- *Q* depends on Φ and
- on word order or the permitted orders of propositional functions

Previous Approach to Grammar Efficiency

- Parsing ratio:

$$Q = Q_o := \frac{s}{a}$$

- s is the number of unambiguous sentences (strings of word classes) permitted in the PoS system
- a is the number of all parsing attempts of all permutations of each sentence in the PoS system (it is assumed that modifiers stand next to their heads)

Turkish PoS System 1

- $l = 4$
- $k = 3$, word classes: V, Λ, M
  (more complicated than the basic types considered above b/c Λ and M overlap)
- Orders of propositional functions:
  RP, rRP, RpP, rRpP
- Sentences: ΛV, ΛΛV – ambiguous , MΛV, ΛMV, ΛΛΛV, MΛΛV, ΛΛMV, MΛMV

$$s = 7$$

Turkish PoS System 2

- Calculating a is complicated:
  $a = 100$, after parsing 32 sentences

$$AE_o = \frac{7}{100} \cdot \frac{4}{3} = \frac{7}{75} = 0.0933$$

- This is low because of the overlapping roles of Λ and M and because of the fixed order of propositional functions

An Example of Parsing Attempts

- ΛΛV → RrP | <u>RpP</u> | <u>rRP</u> | p-
- The approach of "regulated rewriting" is taken.
- Two possible interpretations (underlined) are left. This is why ΛΛV is an ambiguous sentence.
- Other permutations (ΛVΛ and VΛΛ) are parsed in the same way…

New Approach to Grammar Efficiency

The role of a within the parsing ratio is dual:

- It is part of the measure of word-order flexibility/rigidity (all permutations of each possible sentence are considered)
- It also represents indirectly how far the relation Φ is from a bijection (all parsing attempts are considered)

The latter is not related to parsing and can be measured by μ.

The New Formula

$$Q = Q_n := \frac{q}{\mu}$$

- $q = \frac{s}{m}$ — only measures the flexibility of word order, $m = \max\{\hat{s}, f(l)\}$
- $\hat{s}$ is the number of all possible sentences, unambiguous or not
- $f(l)$ is the maximum possible number of orders of propositional functions
- $f(4) = 18, f(3) = 6, f(2) = 2, f(1) = 1$

Turkish PoS System 3

- $m = \max\{32,18\} = 32$ (all 32 possible sentences have to be counted, but they do not have to be parsed)
- $q = \frac{7}{32}, \quad \mu = \frac{3\beta+10}{6} + 1 = \frac{11}{3}$

$$AE_n = \frac{3}{11} \cdot \frac{7}{32} \cdot \frac{4}{3} = 0.0795$$

(cf. $AE_o = 0.0933$)

Relative Grammar Efficiency 1

$$RE = RE(G) = \omega AE$$

- G is the grammar of a PoS system with $|X| = l$ and $|Y| = k$
- A maximally efficient grammar in this class has the greatest value of *AE* and should satisfy certain properties (for instance, it should not permit ambiguity)
- If the maximally efficient grammar exists, its *RE* is set equal to 1

Relative Grammar Efficiency 2

- When the maximally efficient grammar exists and its *AE* is *AE**, then $\omega = \frac{1}{AE^*}$.
  This results in
  $$RE = \frac{Q}{Q^*},$$
  where Q^* is the greatest value of Q for all grammars with $|X| = l$ and $|Y| = k$.
- Otherwise, set ω = 1 and *RE* = *AE*.

Turkish PoS System: Old Approach

- Calculating Q_o^* is also complicated: $Q_o^* = \frac{5}{8}$, after exploring all grammars with all 4 propositional functions and 3 word classes
- Values of Q_o^* are calculated for all k and l in (Vulanović, 2008)
- Relative efficiency of the Turkish PoS system is

$$RE_o = \frac{Q_0}{Q_o^*} = \frac{7}{100} \div \frac{5}{8} = \frac{14}{125} = 0.112$$

Turkish PoS System: New Approach

- Calculating Q_n^* is not so complicated:
  Q_n^*=0.445 (VNMM)

$$RE_n = \frac{Q_n}{Q_n^*} = \frac{3}{11} \cdot \frac{7}{32} \div 0.445 = 0.134$$

(cf. RE_o = 0.112)

Attested PoS System Types

– according to Hengeveld and van Lier (2010).

- This includes systems which are not attested in their “pure” form, but in combination with other types of systems.
- All 5 rigid systems ($k = l, \mu = 1$), VNam, VNaØ, VNØm, VNØØ, and VØØØ, plus 8 flexible PoS systems
- The greatest values of *RE* w.r.t. word order are calculated on the next slide.

Greatest Values of *RE* for Attested PoS System Types

Type	RE_o	RE_n
CCCC	0.286	0.015
VɅɅɅ	0.728	0.797
₱₦₦₱	0.786	0.667
VNMM	0.914	1
V₦₦m	0.800	0.600
V₦₦Ø	1	0.867
$X_3X_3X_3$Ø	1	1
HHØØ	1	1
5 Rigid Types	1	1
Coefficient of Correlation	0.960	

Conclusions

- The new measure is much easier to calculate than the old one.
- The correlation of the old and new values for *RE* is strong for the 13 attested PoS system types.
- It is somewhat weaker when all PoS system types are taken into account: $r = 0.807$.
- Other PoS systems, which (like the Turkish PoS system) are more complicated than the basic ones, can now be approached more easily.

Dziękuję bardzo!