Counting the languages we could speak

Luca La Rocca & Cristina Guardiano

Università di Modena e Reggio Emilia, Italy

www.sce.unimore.it

Scienze della Comunicazione e dell'Economia

Università degli Studi di Modena e Reggio Emilia

Language variation tamed

The Principles & Parameters approach tries to reduce language variability to a finite list of binary options (innately predefined by Universal Grammar and set by language learners on the basis of environmental evidence), such as the one illustrated by the following example:

il mio bel libro	Italian
*mio bel libro	
il libro mio	
*le mon très beau livre	French
mon très beau livre	
*le très beau livre mon	
*the my beautiful book	English
my beautiful book	

* ungrammatical

Three epiphenomenal properties:

1. Co-occurrence of possessives and the article (or other determiners)
2. "Articleless" possessives
3. Postnominal possessives (in languages that have postnominal adjectives)

They co-vary:

	It	Fr	E
1.	yes	no	no
2.	no	yes	yes
3.	yes	no	no‡

‡ no postnominal adjectives in English

They all depend on a unique abstract difference, i.e. the categorization of possessives either as adjectives or as articles (definite determiners):

	It	Fr	E
± D checking poss	+	−	−

This is parameter 48 (out of 51) in Longobardi and Guardiano (in press).

How many possible languages?

A list of n independently set parameters gives 2^n languages: $2^{51} \simeq 2.25 \times 10^{15}$.

Partial interactions between parameters make some languages impossible:

$$s_i = \begin{cases} 0 & \text{if implied by } s_1, \ldots, s_{i-1} \\ \pm 1 & \text{if independently set} \end{cases}$$

for a possible language $s = (s_i)_{i=1}^n$.

Let $\ell_n^i(s_1, \ldots, s_i)$ be the number of valid configurations of n parameters starting with $s_1, \ldots, s_i$: we aim at ℓ_n^0.

In principle, recursive computation is straightforward; in practice, it is only feasible for "small" n (the computation time t_n grows exponentially with n).

Monte Carlo approximation

Let $\sigma^{(1)}, \ldots, \sigma^{(m)}$ be i.i.d. random languages such that

$$\sigma_i^{(1)} = \begin{cases} 0 & \text{if implied by } \sigma_1^{(1)}, \ldots, \sigma_{i-1}^{(1)} \\ \pm 1 & \text{with even odds} \quad \text{otherwise} \end{cases}$$

so that

$$\mathcal{P}\left\{\sigma^{(1)} = s\right\} = 2^{-\|s\|}$$

for any valid configuration s, where $\|s\|$ is the number of nonzero elements in s.

Since, given $\|s\| = k$, all valid s are equiprobable, we approximate ℓ_n^0 by

$$\hat{\ell}_n^0 = \sum_{k=1}^{n} 2^k P_m^k$$

where P_m^k is the proportion of languages with k independently set parameters in $\sigma^{(1)}, \ldots, \sigma^{(m)}$; the corresponding (estimated) standard error is given by

$$SE^2 = \frac{1}{m}\sum_{k=1}^{n} 4^k P_m^k (1 - P_m^k) + \\ - \frac{1}{m}\sum_{h=1}^{n-1}\sum_{k=h+1}^{n} 2^{h+k+1} P_m^h P_m^k$$

Results

Number of possible languages and recursive computation time using the first n parameters in Longobardi and Guardiano (in press); t_{51} was extrapolated via OLS regression ($R^2 = 0.999$) of $\log t_n$ on n.

n	t_n	ℓ_n^0	$\hat{\ell}_n^0 \pm \mathrm{SE}$
15	0.18 s	1570	1571 ± 5
20	1.6 s	12122	12066 ± 54
25	18 s	127184	128409 ± 769
30	3.4 min	1532720	1556308 ± 11962
51	42 days	?	$25.1 \pm 0.5 \times 10^9$

Computations done in R (R Development Core Team, 2008) on an ordinary laptop (MacBook2,1). We let $m = 10^6$. It took 11 minutes to compute $\hat{\ell}_{51}^0$.

Downsizing of grammatical variation due to partial interactions: about 1 every 10^6 parameter configurations is valid (corresponds to a possible language).

Acknowledgements

Advice by Giuseppe Longobardi and Gianpaolo Scalia Tomba is gratefully acknowledged.

References

Longobardi, G. and Guardiano C. (in press). Evidence for syntax as a signal of historical relatedness. *Lingua*, to appear. The electronic database is available at `http://www.units.it/~linglab` (restricted access area).

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.