LINGUIST 168 Introduction to Linguistic Typology

Lecture 3: Methodology, continued
Lexical typology

Ksenia Ershova

April 5, 2021

Plan for today

- Questions about previous lectures, reading, or homework

- Methodology: Language sampling

- Lexical typology

Questions on readings

- What is a probability sample and how can we make any statements about tendencies and universals based on the available set of languages? → will discuss today

- Language documentation versus language description

- Reduplication: full or partial repetition of a segment to express a grammatical meaning

Samoan:

moe 'sleep(singular)'

momoe 'sleep(plural)'

- quote from Bickel 2008 → will discuss today

Questions on readings

- How do linguists decide what aspects of a language to describe?

- How do typologists avoid bias towards certain topics over others in existing descriptions?

- Is language extinction always bad?

In some cases, <u>dormancy </u>is the more appropriate term.

- What can be done to preserve / revitalize endangered languages?

Plan for today

- Questions about previous lectures, reading, or homework

- Methodology: Language sampling

- Lexical typology

Language sampling

- Goal: identify universals, trends and tendencies across all languages

- Problem: we cannot possibly examine all languages

  › ~7000 languages

  › ≈3% of languages that ever existed

  › only ~30% of those are adequately described

- Solution: work with a **sample** of languages

How to assemble a sample

Goal: identify universals, trends, and tendencies across languages which are **independent of genetics, geography, culture and contact**

Roadblocks to a balanced sample

- Many languages are under-described, undocumented or difficult to access

  ➢ Leads to geographically, genetically and culturally biased datasets

E.g. many of the languages in Eurasia have been extensively studied compared to languages of Australia or Africa.

- Language contact often not sufficiently documented

  ➢ Can create false tendencies

E.g. Greek and Bulgarian share many grammatical patterns due to contact.

Activity

In groups, discuss cons of one of the sampling strategies below.

Things to consider

- Size of sample:
  - › the larger the sample, the harder to manage
  - › the smaller the sample, the less representative
- Balance: does this strategy meet our goal of identifying tendencies
  - › independent of geography
  - › independent of genetics
  - › independent of culture
- Difficulties in accurate sampling: recall that 2/3 of languages are under-described in all aspects, including genetic affiliation, culture, and language contact

How to assemble a balanced sample

Strategy 1 (Tomlin 1986):

Each language family in the sample is represented **in proportion to the number of languages in that family.**

E.g. sample consists of 10% of world languages

→ must include 10% of each language family

Pro: accurately represents proportion of a pattern in languages of the world

How to assemble a balanced sample

Strategy 2 (Bybee 1985, Perkins 1989):
Gather languages that are not genetically related amongst themselves (or very distantly related) and are not from the same cultural area.

E.g. one language from the Indo-European language family, one language from the Khoisan language family, one language from the Northeast Caucasian language family, etc.

Pro:

▪ Manageable sample size (~50 languages)

▪ Represents linguistic tendencies that are independent of the spread of a particular language family

What did Bickel mean?

Quote from Bickel 2008 (p.53):

> "In response to this one might choose to admit several languages from each stratum [i.e. genus] in the hope of reducing such effects. However, this option is severely limited because about a third of the proven stocks [i.e. families] in the world are isolates. Since strata [i.e. genera] need to contain the same number of languages, the inclusion of isolates implies that only one datapoint can be admitted for each stratum, even for non-isolates like Romance and Germanic"

- initial strategy: sample with one language from each genus
- problem: might be confounded by language contact / areal influences
- solution: more than one language from each genus
- new problem: isolates (don't have any relatives) will be drowned out in the sample by bigger language families

How to assemble a balanced sample

Strategy 3 (Dryer 1989):

▪ languages are grouped into genera (sing. genus)

Genus = group of related languages that can be traced back to an ancestor ~2500 years ago.

E.g. Romance (Spanish, French, Italian, etc.) and Germanic (English, German, Dutch, Icelandic, etc.)

▪ genera grouped into five large geographic regions

Africa, Eurasia, Australia & New Guinea, North America, and South America

▪ a statistically significant tendency must be observed **in most genera in every geographic region**

Pro: avoids geographic or genetic bias

Genera + geographic region = SOV versus SVO

(51)		Afr	Eura	Austr-NG	NorthAm	SouthAm	Total
	SOV	22	26	19	26	18	111
	SVO	21	19	6	6	5	57

➡ **SOV is a significant cross-linguistic tendency**

Moravscik 2013. Introducing language typology.

Genera + geographic region = SOV versus SVO

	Afr	Eura	Austr-NG	NorthAm	SouthAm	Total
(52)						
SVO	21	19	6	6	5	57
VSO	5	3	0	12	2	22

→ **SVO is a trend, but not a cross-linguistic tendency**

Moravscik 2013. Introducing language typology.

Activity

In groups, discuss cons of one of the sampling strategies below.

Things to consider

- Size of sample:
  - › the larger the sample, the harder to manage
  - › the smaller the sample, the less representative
- Balance: does this strategy meet our goal of identifying tendencies
  - › independent of geography
  - › independent of genetics
  - › independent of culture
- Difficulties in accurate sampling: recall that 2/3 of languages are under-described in all aspects, including genetic affiliation, culture, and language contact