2. Tzv. samplování

2.1. Samplování

Racionále

- paradox: chceme zjišťovat univerzálie lidského jazykového potenciálu, ale nemůžeme zkoumat univerzum jazyků
  – a) ne všechny jazyky aktuálně existují (minulé a budoucí)
  – b) ne všechny existující jazyky jsou (dostatečně) doloženy
  – c) praktická omezení: doložených jazyků je mnoho, ale času a financí málo
- potřebujeme proceduru, která nám umožní provádět INFERENCE o lidském jazykovém potenciálu (= jazyce) bez nutnosti přihlížet ke všem jeho individuálním případům (= jazykům)

Typologický vzorek

- samplování = konstrukce vzorku/samplu
- samplování v lingvistické typologii = konstrukce typologického vzorku jazyků
- typologický vzorek je podmnožinou univerza jazyků

- a) NÁHODNÝ vzorek
- b) PŘÍLEŽITOSTNÍ vzorek *convenience/opportunity sample*
- c) STRATIFIKOVANÝ vzorek

Vhodný typologický vzorek

- A. má umožnit formulaci generalizací různého druhu (variabilita, frekvence, distribuce, preference)
- B. má minimalizovat inferenční chyby
  – musí být reprezentativní vzhledem k univerzu jazyků
  – nesmí být tendenční *biased*; je třeba se vyhnout
  – nedostatečnému zastoupení *underrepresentation*
  – nadměrnému zastoupení *overrepresentation*
- C. nemá odrážet irelevantní faktory
- D. má minimalizovat statistickou/faktickou ZÁVISLOST PŘÍPADŮ (☛ 2.2.)

B. Druhy reprezentativnosti/tendenčnosti vzorku (Bell 1978)

- genealogická[1]
- areální (srov. areály a makroareály; Dryer 1989, Nichols 1992, Dahl 2001)
- sociolingvistická (např. znakové jazyky, kontaktní jazyky, *Ausbausprachen*)
- typologická (zvl. „silné“ typologické parametry)
- bibliografická (zde pouze tendenčnost, nikoli reprezentativnost!)

Příklady tendenčnosti

- běžné nadměrné zastoupení IE a/nebo evropských jazyků z technických a sociálních důvodů: dostupnost mluvčích, deskriptivní tradice aj.
- slovosled AN preferován v sev. Euroasii (IE, uralské, „altajské“, čínština); nadměrné zastoupení euroasijských jazyků ve starších vzorcích vedlo k formulaci neplatné generalizace: pokud OV, pak AN (srov. Dryer 1989)

Příklad konfliktu mezi různými druhy reprezentativnosti

- genealogické izoláty a malé rodiny často součástí jaz. areálu
- např. Papua-Nová Guinea: mnoho genealogických skupin v malé oblasti

C. Příklady faktorů, které nesouvisejí se strukturou jazyka

- počet zemí, v nichž se jazykem mluví
- počet jazyků v genealogické skupině jazyků
- počet jazyků v makroareálu
- počet mluvčích jazyka (ale: protipříklady!)

[1] Pozor na cirkularitu: některé „genealogické“ skupiny (např. tzv. uralo-altajská) jsou ustanoveny na základě typologických shod.

Typologie bez samplování?

- i studie bez samplovací procedury mohou být plodné: „připravují půdu“, mohou odkrývat základní trendy
- ale: výsledky musejí být ověřeny pomocí samplování!

2.2. Teoretické problémy

Předpoklad reprezentativnosti

- = existující, příp. zdokumentované jazyky reprezentují lidský jazykový potenciál
- předpoklad je problematický: pokud platí aspoň jedna z následujících hypotéz, empirický výzkum univerzálií necharakterizuje lidský jazykový potenciál
- a) všechny jazyky světa můžou být genealogicky příbuzné, tzv. *Proto-World* (Comrie 1981)
- b) všechny jazyky světa můžou tvořit jediný globální jazykový areál (Dryer 1989)
- c) existující, příp. zdokumentované jazyky se zachovaly díky historické náhodě
  – Chomského anekdota: kdyby všechny jazyky kromě jednoho vymřely, budeme považovat všechny rysy tohoto jazyka za univerzální?
  – existující jazyky můžou být *lucky rather than natural*, mohly se zachovat díky technologické a/nebo politické dominanci svých mluvčích (Maddieson 1999)

Příklad

- základní slovosled OS je velmi řídký, doložený jen v několika (ohrožených) jazycích
- kdyby typologové začali zkoumat slovosledné univerzálie až po zániku těchto jazyků, formulovali by substanční univerzálii: každý jazyk má základní slovosled SO

D. Statistická (ne)závislost případů (Perkins 1989)

- < tzv. Galtonova námitka v antropologii: kulturní typy můžou být „závislé“ kvůli migraci nositelů nebo difuzi prvků
- asociace = absence statistické nezávislosti
- požadavek: statistická nezávislost relevantních vlastností jazyků ve vzorku
  – typologický parametr & genealogická afiliace
  – typologický parametr & areální afiliace
  – typologické parametry P & Q

Příklad 1

- typologický parametr rozlišuje tři typy: T1, T2, T3
- svět se dělí na tři makroareály: A1, A2, A3
- všechny jazyky A1 jsou T1, A2 jsou T2 a A3 jsou T3 (absolutní statistická korelace typů s makroareály)
- ke správné generalizaci stačí vzorek o 3 jazycích (po jednom z A1, A2 a A3)

Příklad 2

- Dryer 1989: chtěl zabránit typologické tendenčnosti, a proto chtěl mít ve vzorku jazyky různých typů podle parametru OV/VO
- Perkins 1989: Dryer ale neprokázal, že chtít mít ve vzorku různé typy podle P (= OV/VO) statisticky neovlivňuje výsledky zkoumaného jevu Q

- univerzální vzorek je nežádoucí, nadhodnocuje rysy velkých jaz. rodin, areálů s mnoha jazyky apod.
- požadavek statistické nezávislosti je v KONFLIKTU s požadavkem reprezentativnosti (kde: univerzální vzorek naopak žádoucí)

- konflikt je neřešitelný: vzhledem k existenci makroareálů můžeme získat jen 10 statisticky nezávislých případů, ty pak nejsou reprezentativní (Dryer 1989)

- ale: statistická závislost neimplikuje FAKTICKOU závislost případů (Croft 2003, Dryer 1989, Comrie 1993)
  – i jevy v genealogicky nebo areálně zpřízněných jazycích můžou být historicky nezávislé
  – např. všechny odlišnosti blízce příbuzných jazyků (tj. jejich inovace), jsou fakticky nezávislé (Comrie 1993)

2.3. Druhy typologických vzorků

Velikost vzorku
- a) konstrukce *top-down*
  – nejdříve určíme počet jazyků ve vzorku, pak vzorek stratifikujeme a vybíráme konkrétní jazyky
  – problém: velikost minimálního vzorku závisí na zvoleném typologickém parametru, není dán apriori (Perkins 1989)
- b) konstrukce *bottom-up*
  – nejdříve vybíráme konkrétní jazyky, tak dospějeme ke vzorku urč. velikosti

Druhy vzorků podle cíle
- a) vzorek VARIANTNOSTI *variety sample*
  – cíl: najít všechny různé realizace urč. jevu (hledáme, dokud nenajdeme nový nezávislý případ, příp. do vyčerpání logických možností)
  – důraz na reprezentativnost
- b) vzorek PRAVDĚPODOBNOSTI *probability sample*
  – cíl: identifikovat typologické kovariace (frekvence a preference)
  – důraz na nezávislost případů

Několikeré samplování
- a) pilotní vzorek pro zjištění možných typů vybraného parametru
- b) větší (stratifikovaný) vzorek variantnosti pro zjištění distribuce typů
- c) menší (statisticky testovaný) vzorek pravděpodobnosti pro formulaci signifikantních generalizací

Druhy vzorků podle struktury
- a) vzorek PROPORCIONÁLNÍ
  – každá skupina jazyků (zvl. genealogická) má rovnou reprezentaci
  – problém: postihuje frekvenci v jazycích, nikoli preferenci lidského jazykového potenciálu
- b) vzorek HIERARCHICKÝ
  – pracuje s několika hierarchizovanými úrovněmi klasifikace jazyků (příklady ☛ 2.4.)

Příklad 1
- proporcionální vzorek: Bell 1978
- definoval genealogickou skupinu arbitrární časovou hloubkou 3500 let
- dospěl k 478 skupinám (např. IE má 12 skupin)
- vzorek o méně než 478 jazycích je nutně nereprezentabilní

Příklad 2
- předpokládejme, že na světě existuje 1000 jazyků 11 genealogických skupin (900 + 10 + 10 + 10 + 10 + 10 + 10 + 10 + 10 + 10 + 10)

- jazyky největší skupiny mají slovosled SVO, jazyky všech ostatních skupin mají slovosled SOV
- poměr jazyků: 900 SVO :: 100 SOV, tj. frekventovanější je SVO (90%)
- poměr gen. skupin: 1 SVO :: 10 SOV, tj. „preferovanější" je SOV (91%)

2.4. Dva „úspěšné" vzorky

Dryer 1989

- procedura kontroluje tendenčnost na dvou úrovních (hierarchický vzorek)
- a) na úrovni makroareálů
- b) na úrovni tzv. gener: GENUS = genealogická jednotka s časovou hloubkou cca 3000-4000 let (~ větev IE jazyků)
  – pokud je genus typologicky jednotný > jeden bod
  – pokud je v rámci genera více typů > více bodů
- umožňuje využít všech dostupných dat
- problémy
  – jak vybírat jazyky v rámci genera?
  – konkrétní vymezení makroareálů (Nichols 1992)
  – i v čas. hloubce, která definuje dnešní genera, je přítomna tendenčnost díky tehdejší existenci větších a menších gener (Croft 1995)

Rijkhoff et al. 1993

- procedura kontroluje pouze GENEALOGICKOU tendenčnost na dvou úrovních (hierarchický vzorek)
- a) mezi tzv. FYLY: každé, včetně izolátů, zastoupeno (27 podle Ruhlena 1987)
- b) v rámci tzv. fyl: bere v úvahu jejich interní diverzitu (= vnitřní různorodost)
- tzv. HODNOTA DIVERZITY *diversity value*
  – reprezentuje ne časovou hloubku, nýbrž vnitřní strukturu genealogického „stromu"
  – vychází z počtu „štěpení" mezi fylem a jednotlivým jazykem
  – významnější jsou časově hlubší „štěpení", je k dispozici delší doba na vytvoření odlišností
- PROBLÉMY
  – modely interní diverzity různých genealogických skupin založeny na různých kritériích (různí autoři, různé metody)
  – interní diverzita někt. genealogických skupin není dostatečně popsána; zvl. u méně popsaných skupin může být větší, než se předpokládá
  – sám pojem genealogického „stromu" (☛ Úkol II.2)

Úkol II.1

- Vytvořte typologický vzorek
  1. o 20 jazycích;
  2. hierarchický: vezměte v úvahu a) kontinentální areály a b) genealogickou klasifikaci včetně interní diverzity jednotlivých genealogických skupin;
  3. předpokládejte situaci, kdy na světě existují pouze tři genealogické skupiny jazyků: alžské jazyky *Algic*, indoevropské jazyky a izolát buruština *Burushaski* (vycházejte z genealogické klasifikace na serveru *Ethnologue*);[2] http://www.ethnologue.com/family_index.asp
  4. teoreticky svůj vzorek a jeho konstrukci popište.

[2] ☛ http://en.wikipedia.org/wiki/Ethnologue

Úkol II.2

- Je jazyk jakožto jednotka typologického vzorku jasně DEFINOVATELNÝM OBJEKTEM?
- Proč je pojem GENEALOGICKÉHO „STROMU“ problematický pro modelování interní diverzity jazykových skupin?

Literatura: základní práce o samplování

Bell, Alan. 1978. Language samples. In: Greenberg, Joseph H., Charles A. Ferguson & Edith A. Moravcsik (eds.) *Universals of human language*, Vol. 1. 123-156. Stanford: Stanford University Press.

Comrie, Bernard. 1993. Language universals and linguistic typology: data-bases and explanations. *Sprachtypologie und Universalienforschung* 46: 3-14.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13, 257-292.

Perkins, Revere D. 1989. Statistical techniques for determining language sample size. *Studies in Language* 13, 293-315.

Rijkhoff, Jan & Dik Bakker. 1998. Language sampling. *Linguistic Typology* 2, 262-314.

Rijkhoff, Jan, Dik Bakker, Kees Hengeveld, & Peter Kahrel. 1993. A method of language sampling. *Studies in Language* 17, 169-203.

Další citovaná literatura

Comrie, Bernard. 1981, 1989[2]. *Language universals and linguistic typology: Syntax and morphology*. Oxford: Blackwell.

Croft, William. 1990, 2003[2]. *Typology and universals*. Cambridge: Cambridge University Press.

Dahl, Östen. 2001. Principles of areal typology. In: Haspelmath, Martin, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.) *Language typology and language universals: an international handbook*, Vol. 2. 1456-70. Berlin: Mouton de Gruyter.

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.