4. Korpusová lingvistika III.
Case study - morfologická produktivita

---

1) Co je produktivita?
2) Baayenův přístup založený na korpusu
3) Hlavní pojmy
4) Předpoklady a problémy
5) Aplikace
6) Citovaná literatura

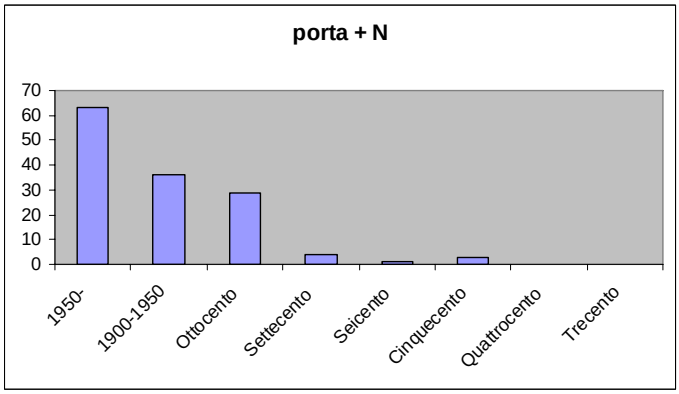
Ad 1) Co je produktivita?

- D. Corbin 1987, 177: "... la productivité désigne en fait à la fois la régularité des produits de la règle, **la disponibilité** de l'affixe, c'est-à-dire précisément la possibilité de construire des dérivés non attestés, de combler les lacunes du lexique attesté, et **la rentabilité**, c'est-à-dire la possibilité de s'appliquer à un grand nombre de bases et/ou de produire un grand nombre de dérivés attestés."
- L. Bauer 2001: *availability* / *profitability*
- Disponibilita = prostá přítomnost prostředku v systému → KVALITATIVNÍ PŘÍSTUP
- Rentabilita = malá / velká "výnosnost" daného prostředku → **KVANTITATIVNÍ PŘÍSTUP**

KVANTITATIVNÍ PŘÍSTUP

a) dictionary-based b) corpus-based

Ad a) Produktivita definovaná jako *type frequency* (*porta* + *subst.* podle slovníku DISC 2004)



Ad b) V korpusu je kromě *type frequency* také *token frequency*, tj. v korpusu je nejen nějaký počet lemmat / typů *porta* + *subst.*, ale také nějaký počet *výskytů* těchto jednotlivých lemmat.

Např. *porta* + *subst.* vypadají v korpusu *La Repubblica* takto (Ricca 2008):

slovotvorný prostředek	počet lemmat (V)	výskyty (N)	velikost korpusu (F)
porta + *subst.*	214	46289	~ 330 mil.

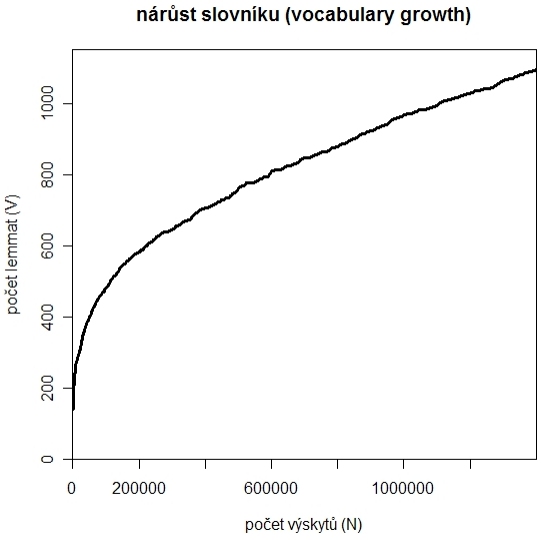
Nebo slovesa s prefixem *ri-* v témže korpusu (Baroni 2007; Baroni, Evert 2006)

slovotvorný prostředek	počet lemmat (V)	výskyty (N)	velikost korpusu (F)
ri + sloveso	1098	1399898	~ 330 mil.

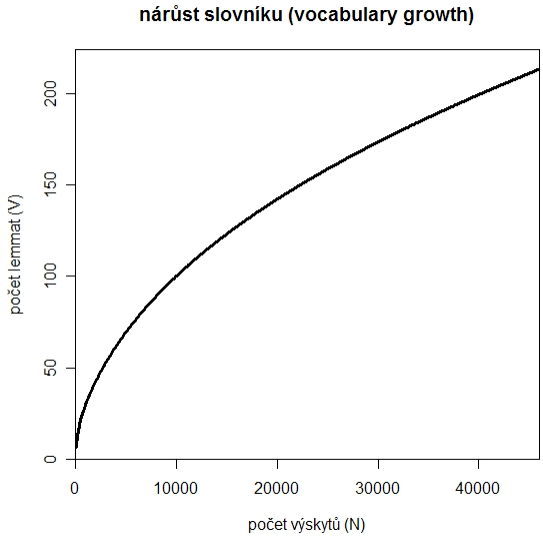
Ad 2) Baayenův kvantitativní přístup

- Baayen 1992, 113: Vztah mezi *token frequency* (N) a *type frequency* (V) může být nahlédnut jako funkce V(N): počet V je funkcí N - se vzrůstající hodnotou N poroste i V; N je dáno velikostí korpusu. Tento vztah je vidět na "křivce nárůstu slovníku" (*vocabulary growth curve*)

Slovesa s prefixem *ri-* (empirická křivka)



Kompozita *porta + subst.* (interpolovaná křivka)



- Baayen zároveň definuje i *tempo*, jakým slovník narůstá (tzv. *vocabulary growth rate*). Rychlost nárůstu lze určit ze *sklonu* křivky v daném bodě, který se vyjádří poměrem všech *hapax legomena*, tj. lemmat, která se objevují v korpusu jen jednou, a jejich celkovým počtem výskytů, tedy $P = V_1 / N$

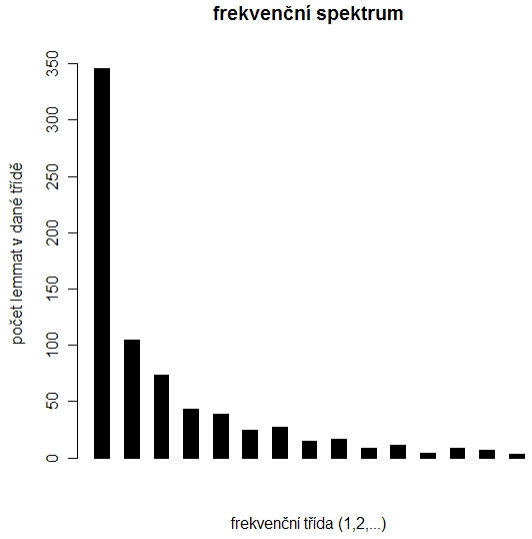
slovotvorný prostředek	počet lemmat (V)	výskyty (N)	počet *hapax legomena* (V_1)	P (V_1/N)
ri + sloveso	1098	1399898	346	0,00024
porta + substantivum	214	46289	103	0,00222

- Rovnice $P = V_1 / N$ je vlastně podobná té, která je ukryta pod MI-score, tj. pravděpodobnost nalezení daného slova uvnitř korpusu f(x) / N; zde jde o pravděpodobnost nalezení nového lemmatu uvnitř korpusu; typy, které se vyskytují jen jednou, jsou vyděleny všemi ostatními výskyty.
- "The growth rate is a probability, the probability that, after having read *N* tokens, the next token sampled represents an unseen type, a word that did not occur among the preceding *N* tokens." (Baayen 2008, s. 222)
- Baayen tedy vidí produktivitu (ve smyslu *rentability*) jako pravděpodobnost toho, že narazíme na nové slovo, když budeme postupně procházet daný korpus (popř. to lze interpretovat jako nějakou časovou dimenzi)
- Produktivnější procesy - ty, které rostou rychleji, mají vyšší P hodnotu.

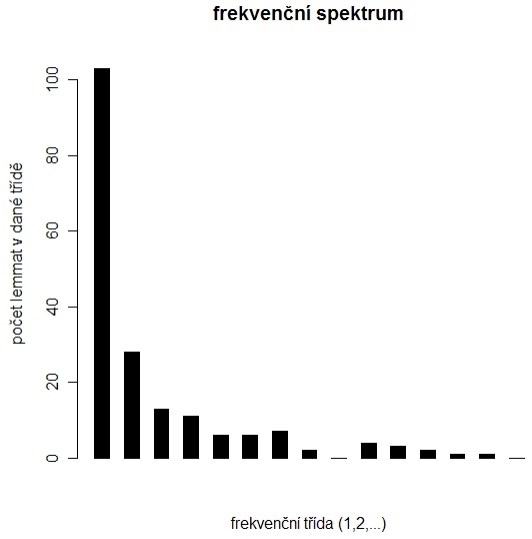
3) Hlavní pojmy

- Tento přístup je dnes zcela běžný (a poměrně oblíbený) a spadá do toho, čemu se říká lexikální statistika (a také je za tím spousta složité matematiky ...)
- Stefan Evert a Marco Baroni - prográmek *zipfR* (http://zipfr.r-forge.r-project.org/), "package" k použití ve statistickém softwaru R (www.r-project.org) (úvod do R pro lingvisty Baayen 2008) (Baroni - Evert 2006; Evert - Baroni 2006b, 2007)
- Předpoklad: produktivní procesy budou *růst*, neproduktivní by se měly nějak blížit nule. Produktivní procesy vykazují typické **frekvenční spektrum / frekvenční profil** (Baroni, Distributions in text, čteme ...) (jakkoli toto spektrum je charakteristické pro všechny korpusy *en bloc*; vybereme-li určité - extrémní - kategorie, měl by tu být rozdíl - bude lépe vidět na růstových křivkách viz níže).

Frekvenční spektrum sloves s prefixem *ri-*



Frekvenční spektrum kompozit *porta + subst.*



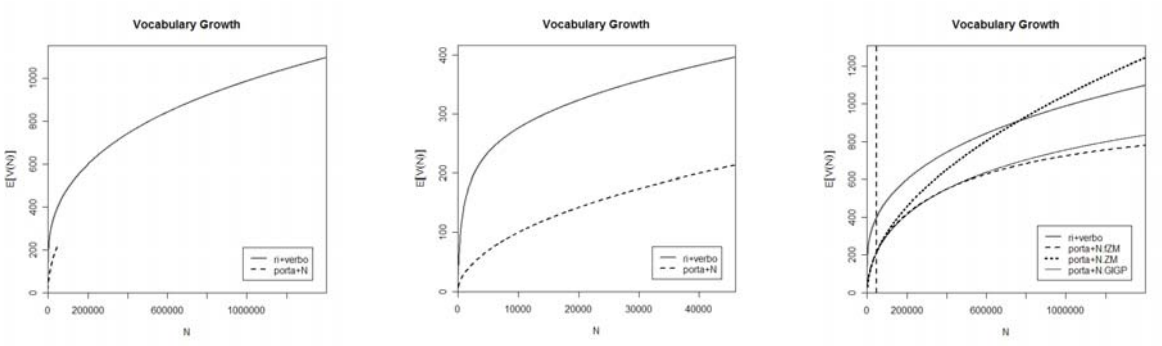
- Má-li daná kategorie takovéto spektrum, bude mít i křivku nárůstu velmi podobnou těm, které jsme viděli:

o **EMPIRICKÁ KŘIVKA** = zachycuje hodnoty postupně "naměřené" /pozorované/ v jednotlivých bodech, např. po 1000 výskytech mám x lemmat, y hapaxů, po 2000 výskytech mám x + n lemmat, y + n hapaxů atd. viz např. prefix ri-:

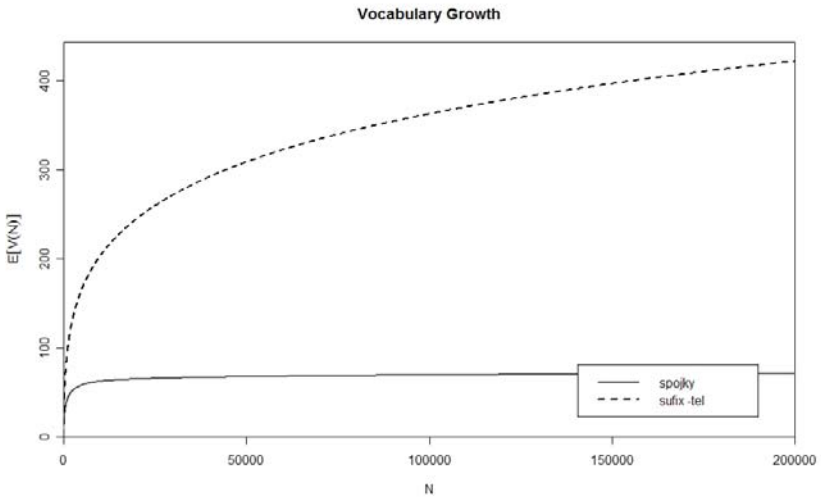
N	*V*	V_1
200000	583	176
400000	707	213
600000	810	259
800000	881	274
1000000	969	306
1200000	1029	314

o **INTERPOLOVANÁ KŘIVKA** = obvykle je interpolace pospojování bodů na nějakém grafu tak, abychom viděli nějakou souvislou tendenci, a tak, abychom naznačili nějaké "skryté hodnoty pro případy, které nebyly změřeny" (srov. Volín, J. *Statistické metody ve fonetickém výzkumu*, s. 41).

- Interpolace empirických křivek je zde realizována pomocí tzv. *binomické interpolace* (binomial interpolation), což už není jen grafické pospojování, ale matematická operace, která "produkuje očekáváné (predikované, *expected values*) hodnoty V, V_1 (atd.) pro různé hodnoty N" (popis u Baayena 2001; Baroni - Evert 2006 aj.)
- **INTERPOLACE** produkuje tyto hodnoty až do empirického N, zatímco ...
- **EXTRAPOLACE** produkuje tyto hodnoty pro jakoukoli velikost N. Extrapolace je ale složitá věc, protože jde za reálná data (Evert, Baroni 2006a); vychází ze speciálních statistických modelů frekvenční distribuce, tj. "rozdělení" specifické pro frekvence slov. Existují tři dobré modely (GIGP, ZM, fZM; Baayen 2001; Evert 2004), založené na tom, čemu se říká LNRE (Large Number of Rare Events; speciální případ rozdělení vzácných jevů), všechny jsou implementovány v zipfRu.
- **INTERPOLOVANÁ / EXTRAPOLOVANÁ KŘIVKA** slouží k tomu, abychom mohli porovnat kvantitativně různé procesy, např. právě ri+sloveso a porta+subst. (nebo nějaké zajímavější dvojice...), a to buď tak, že "sestoupíme" s jedním procesem na hodnotu toho druhého (interpolace), popř. "vystoupíme" či "natáhneme" hodnoty jednoho procesu na hodnoty nějakého "většího". Jinak to nepůjde ... to je vidět:



- **JAK VYPADAJÍ KŘIVKY PRODUKTIVNÍHO VS. NEPRODUKTIVNÍHO PROCESU?**
- Teoreticky tak, že u neproduktivního procesu nebude už růst hodnota V, ale bude narůstat jen frekvence výskytů (Baroni - Evert ukazují rozdíl na it. zájmenech vs. prefix ri-).
- Předpokladem je důkladný "pre-processing", tj. poctivě lemmatizované seznamy daných kategorií (preprocessing - v SYN2000 lemma + manuální korekce/eliminace (viz níže příklady)
- Příklad: na jedné straně spojky (jako neproduktivní), na straně druhé derivovaná substantiva na sufix *-tel* (učitel, zhotovitel atd.) (jako produktivní)



lemmata	počet lemmat (V)	výskyty (N)	počet *hapax legomena* (V_1)
sufix -tel	453	278226	97
spojky	76	6967738	0

4) Předpoklady a problémy

- Jak už jsme viděli, srovnávat lze jen se stejnou hodnotou N; takže je třeba použít interpolaci/extrapolaci pro sjednocení N, pokud pracujeme s různě velkými korpusy nebo různě "rentabilními" prostředky (nebo použít metodu tzv. *variable-corpus approach* - Gaeta - Ricca 2003; 2006)
- Je třeba ze seznamu lemmat eliminovat některé formace, které nelze považovat za morfologicky komplexní jednotky (jak by řekla Corbinová "mots complexes construits") - a které nám tam naopak automatické vyhledávání vpustí, např. určitě patří *činitel, učitel, věřitel*, ale ne *datel*, *přítel* a mnoho dalších /tj. zlikvidovat tzv. *extraction noise*/.
- A velký problém - co je hapax? Je potřeba se ujistit, že všechny hapaxy jsou neologismy? (K tomu mnoho diskuse: Dal 2003, Plag 1999, 2006 atd. atd.), protože jen tak můžeme ty křivky považovat za věrný obrázek pravděpodobnosti, že narazíme na **nově utvořené slovo.**
- Hapax = neologismus - těžko prokazatelné, složitá práce; různé výsledky podle korpusů apod.
- Jaký je rozdíl mezi slovy s vysokou frekvencí a s nízkou frekvencí?
- Plag 2006: "[such] words are crucial for the determination of the productivity of a morphological process because in very large corpora *hapax*es tend to be words that are unlikely to be familiar to the hearer or reader. Complex unknown words can be understood at least in those cases where an available word-formation rule allows the decomposition of the newly encountered word into its constituent morphemes and thus the computation of the meaning on the basis of the meaning of the parts. The word-formation rule in the mental lexicon guarantees that even complex words with extremely low frequency can be understood. Thus, with regard to productive processes, we expect large numbers of low frequency words and small numbers of high frequency words, with the former keeping the rule alive. In contrast, unproductive morphological categories will be characterized by a preponderance of words with rather high frequencies and by a small number of words with low frequencies."
- Mrkněme na česká slova na -tel (vysoké frekvence vs. *hapaxy*)

LEMMA	FREKVENCE	LEMMA	FREKVENCE
ředitel	42694	využívatel	1
obyvatel	21288	slibovatel	1
představitel	19612	zahubitel	1
přítel	18038	ponižovatel	1
majitel	17816	účtovatel	1
podnikatel	15318	zpytatel	1
učitel	10022	usměrňovatel	1
spisovatel	8411	uskutečnitel	1
uživatel	7703	zvěčňovatel	1
velitel	6710	ovlivňovatel	1
nepřítel	5937	vzkřísitel	1
pachatel	5673	pozměňovatel	1

- Mrkněme naopak na spojky:

LEMMA	FREKVENCE	LEMMA	FREKVENCE
a	2831331	ledaže	281
že	867283	ježto	133
i	543845	jakože	93
ale	389016	sotvaže	78
jako	319539	liž	47
aby	205966	pakli	33
nebo	201500	anžto	30
když	193081	zdaliž	18
však	190568	děleno	18
až	162687	ačli	13
než	153767	zdaž	5
ani	126062	pokaď	2

5) Aplikace

1) Srovnání slovotvorných prostředků v jednom jazyce (např. Baayen - Renouf 1996; Gaeta - Ricca 2002; 2003; 2006; Baroni 2007; Dal 2003; Dal et al. 2007 atd.)
2) Diachronie - srovnání vývoje produktivity v nějakých časových úsecích (t_1, t_2, ...) (např. Lüdeling - Evert 2005; Štichauer 2009; moc studií není; diachronie je problematická - nedostatek rozsáhlých korpusů a jejich heterogenní charakter)
3) Cross-linguistic studie ... (např. italské *-tore*, fr. *-eur* vs. anglické *-er*?)

6) Citovaná literatura

BAAYEN, Harald (1992), Quantitative aspects of morphological productivity, in BOOIJ, G., MARLE, J. VAN (eds.). *Yearbook of Morphology 1991*, Dordrecht, Kluwer, p. 109-149.
BAAYEN, Harald (2001), *Word frequency distributions*, Dordrecht, Kluwer.
BAAYEN, Harald (2008), *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*, Cambridge, Cambridge University Press.
BAAYEN, Harald, RENOUF, Antoinette (1996), Chronicling *the Times*. Productive Lexical Innovations in an English Newspaper, *Language*, 72, p. 69-96.
BARONI, Marco (2007), I sensi di *ri-*. Un'indagine preliminare, in MASCHI, R., PENELLO, N., RIZZOLATTI, P. (eds.), Miscellanea di studi linguistici offerti a Laura Vanelli, Udine, Forum, p. 163-171.
BARONI, Marco (to appear 2009). Distributions in text, in LÜDELING, Anke, MERJA, Kytö (eds.), *Corpus Linguistics: An International Handbook*, Berlin, Mouton de Gruyter..
BARONI, Marco, EVERT, Stefan (2006), The *zipfR* package for lexical statistics: A tutorial introduction. Disponibile su : http://www.cogsci.uni-osnabrueck.de/~severt/zipfR/.
BAUER, Laurie (2001), *Morphological Productivity*, Cambridge, Cambridge University Press.
CORBIN, Danielle (1987), *Morphologie dérivationnelle et structuration du lexique*, 2 voll., Tübingen, Niemeyer.
DAL, Georgette (2003), Productivité morphologique: définitions et notions connexes, *Langue française*, 140, p. 3-23.

DAL, GEORGETTE - FRADIN, B. - GRABAR, N. - LIGNON, S. - NAMER, F. - PLANCQ, C. - YVON, F. - ZWEIGENBAUM, P. (2007), Linguistic prerequisites to the calculation of morphological productivity and first results. Talk presented at *Journées ATALA*, Paris, November 10, 2007.
DISC - *Dizionario Italiano Sabatini-Coletti* Compact versione 1.1. Milano, Giunti, 1997.
EVERT, Stefan (2004), A simple LNRE model for random character sequences, *Proceedings of JADT 2004*, p. 411-422.
EVERT, Stefan, BARONI, Marco (2006a), Testing the extrapolation quality of word frequency models. Proceedings of Corpus Linguistics 2005, http://www.corpus.bham.ac.uk/PCLC/.
EVERT, Stefan, BARONI, Marco (2006b), The *zipfR* library: Words and other rare events in R. Presentation at *useR! 2006: The Second R User Conference*, Vienna, Austria.
EVERT, Stefan, BARONI, Marco (2007), *zipfR*: Word frequency distributions in R, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Session*, Prague, Czech Republic.
GAETA, Livio, RICCA, Davide (2002), Corpora testuali e produttività morfologica: i nomi d'azione in due annate della *Stampa*, in BAUER, R. - GOEBL, H. (a cura di). *Parallela IX. Testo – variazione – informatica. Text – Variation – Informatik*, Wilhelmsfeld, Gottfried Egert Verlag, p. 223-249.
GAETA, Livio, RICCA, Davide (2003), Frequency and productivity in Italian derivation: A comparison between corpus-based and lexicographical data, *Italian Journal of Linguistics / Rivista di Linguistica* 15, 1, p. 63-98.
GAETA, Livio, RICCA, Davide (2006), Productivity in Italian word formation: A variable-corpus approach, *Linguistics* 44, 1, p. 57-89.
LÜDELING, ANKE. - EVERT, STEFAN. The Emergence of Non-Medical -itis. Corpus Evidence and Qualitative Analysis. In Kepser, S. - Reis, M. (eds.). *Linguistic evidence. Empirical, Theoretical, and Computational Perspectives*. Berlin: Mouton de Gruyter, 2005, 315-333.
PLAG, Ingo (1999), *Morphological Productivity. Structural Constraints in English Derivation*, Berlin, Mouton de Gruyter.
PLAG, Ingo (2006), Productivity, in AARTS, B., MCMAHON, A, *The Handbook of English Linguistics*, Oxford, Blackwell, p. 537-557.
RICCA, Davide (2008), VN compounds in Italian: Data from corpora and theoretical issues. Comunicazione presentata al convegno CompoNet Congress on Compounding, Bologna, 6-7 giugno 2008.
ŠTICHAUER, PAVEL (2009). Morphological productivity in diachrony: the case of the deverbal nouns in *-mento*, *-zione* and *-gione* in Old Italian from the 13th to the 16th century. In *Selected Proceedings of the 6th Décembrettes*, ed. Fabio Montermini, Gilles Boyé, and Jesse Tseng, 138-147. Somerville, MA: Cascadilla Proceedings Project, 2009. Dostupné z http://www.lingref.com/cpp/decemb/6/abstract2241.html