Udo Strauss
Fengxiang Fan
Gabriel Altmann

Edice Qfwfq

Kvantitativní lingvistika

Vybrané problémy I

Olomouc
2014

Překlad:
Miroslav Kubát, Radek Čech

Přeloženo podle:
Udo Strauss, Fengxiang Fan, Gabriel Altmann:
Problems in quantitative linguistics 1.

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid

Tato publikace vychází v rámci grantu Inovace studia obecné jazykovědy a teorie komunikace ve spolupráci s přírodními vědami. reg. č. CZ.1.07/2.2.00/28.0076.

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

1. vydání

ISBN 978-80-244-4350-8

Obsah

Předmluva **9**
Poděkování **15**
1 Fonologie a písmo **16**
1.1 Přízvuk a frekvence 16
1.2 Kanonická struktura slova 17
1.3 Konsonanty a shluky 17
1.4 Distribuce kanonických tvarů 18
1.5 Distribuční počet 18
1.6 Distribuční mezery 20
1.7 Vývoj komplexity písma 21
1.8 Využití kanonických tvarů 22
1.9 Frekvence písmen 22
1.10 Měření distinkce 23
1.11 Měření ornamentality 24
1.12 Frekvence fonémů a slov 25
1.13 Inventář fonémů a délka slova 26
1.14 Mocninný zákon 27
1.15 Pořadí typů slabik 28
1.16 Komplexita písma 28
1.17 Zjednodušování písma 29
1.18 Frekvence slabik 30
1.19 Struktura slabik 31
1.20 Tendence k vokálové harmonii 34
1.21 Dvojdimenzionální struktura slabik 35
1.22 Délka slova a suprasegmentální jednotky 36
2 Gramatika **38**
2.1 Behagelův zákon 38
2.2 Spoluvýskyt a koheze 40
2.3 Kotextualita a variace 41

2.4 Frekvence a pád 41
2.5 Frekvence a koheze 42
2.6 Frekvence a derivace 43
2.7 Frekvence a nepravidelnost 44
2.8 Frekvence a valence 46
2.9 Frekvence větných vzorců 47
2.10 Gramatikalizace 47
2.11 Frekvence morfémů 48
2.12 Polysémie morfému a jeho frekvence 49
2.13 Morfologická produktivita kmenů 49
2.14 Sekvenční frekvence slovních tříd 50
2.15 Klasifikace sloves 51
2.16 slovesa a osoby 52
2.17 Distribuce slovních tříd 55
3 Kompozita a lexikologie 56
3.1 Stáří slova a tendence k tvoření kompozit 56
3.2 Kolokace 57
3.3 Délka kompozita a délka komponentu 58
3.4 Délka kompozit a jejich kotextualita 59
3.5 Délka kompozit a polysémie 59
3.6 Délka kompozit a sémantická shoda 60
3.7 Kompozita a sémantická shoda 61
3.8 Skládání slov a asociace 62
3.9 Skládání slov a emocionalita 63
3.10 Kotextualita a tendence k tvoření kompozit 63
3.11 Disortativita skládání slov 64
3.12 Distribuce délky kompozit 65
3.13 Distribuce synonym 66
3.14 Nárůst počtu přejatých slov 67
3.15 Lexikální řetězce 68
3.16 Lexikální sítě 70
3.17 Délka kmenu a tendence k tvoření kompozit 71

3.18 Délka slov a synonymie 72
4 Textologie 74
4.1 Asociační graf textu 74
4.2 Shlukování autosémantik v daných intervalech 75
4.3 Carrollův vektor 76
4.4 AlternatIvnÍ type-token index 77
4.5 Kotextualita a frekvence 78
4.6 Vzdálenosti mezi stejně dlouhými větami 79
4.7 Vzdálenosti mezi lexémy 79
4.8 Eufonie 80
4.9 Hirschův-Popescův bod 82
4.10 Hreby 83
4.11 Hurstův exponent 85
4.12 Köhlerovy motivy délek slov 1 87
4.13 Köhlerovy motivy délek slov 2 89
4.14 Köhlerovy motivy délek slov 3 89
4.15 Köhlerovy motivy délek vět 90
4.16 Lorenzova křivka 91
4.17 Ljapunovův koeficient 92
4.18 Minkowského klobása 93
4.19 N-gramy a motivy délek 94
4.20 Nominální styl 96
4.21 Fonetická agregace 97
4.22 Analýza polylogu 98
4.23 POPESCŮV INDEX SLOVNÍHO BOHATSTVÍ 100
4.24 Indexy 101
4.25 Rytmické jednotky 102
4.26 Obtížnost textu 104
4.27 Tematická koncentrace 105
4.28 Tokeny a Ljapunovův koeficient 106
4.29 Vztah typů a tokenů 107
4.30 Slovesný profil 108

4.31 Bohatství slovníku a reference 111
4.32 Frekvence slov 1 112
4.33 Frekvence slov 2 113
4.34 Frekvence slov 3 114
5 Frekvence a délka 115
5.1 Distribuce délky slov 1 115
5.2 Distribuce délky slov 2 116
5.3 Distribuce délky slov a Ordovo kritérium 117
5.4 Frekvence a tendence k tvoření kompozit 117
5.5 Frekvence a užitečnost písmen 118
5.6 Frekvence a příznakovost/komplexita 119
5.7 Frekvence a slovosled ve frázích 122
5.8 Frekvence a komplexita fonému 123
5.9 Frekvence a forma fonému 124
5.10 Frekvence a produkční úsilí 125
5.11 Frekvence a produktivita 126
5.12 Frekvence a redukce 126
5.13 Frekvence a rozmanitost 129
5.14 Délka a frekvence 130
5.15 Délka a polysémie 133
5.16 Délka a slovní druhy 1 135
5.17 Délka a slovní druhy 2 136
5.18 Délka věty a délka klauze 136
5.19 Délka slova a polytextualita 138
5.20 Délka slov a pozice ve větě 139
5.21 Délka slova/morfému a kompozita 141
6 Sémantika, synergetika a psycholingvistika 142
6.1 Abstraktnost 142
6.2 Distribuce polysémie 143
6.3 Obeznámenost a frekvence 144
6.4 Obeznámenost se slangovými slovy 145
6.5 Frekvence kanji 146

6.6 Učení a komplexita 147
6.7 Učení se u dětí 148
6.8 Význam a frekvence 149
6.9 Inventář morfémů a jejich polysémie 150
6.10 Morfologie vs. fonologie 151
6.11 Inventář fonémů vs. délka morfémů 152
6.12 Polysémie a skládání slov 153
6.13 Sémantické třídy 153
6.14 Sémantická diverzifikace 154
7 Typologie 156
7.1 Entropie a syntetismus 156
7.2 Homonymie a synonymie afixů 1 158
7.3 Homonymie a synonymie afixů 2 159
7.4 Flexe obecně 160
7.5 Délka morfu 161
7.6 Popescův typologický indikátor a 162
7.7 Délka kořenu a rozsah derivace 164
7.8 Syntetismus v jazyce 165
7.9 Vokalický jazyk 166
7.10 Délka slova a kongruence 167
7.11 Pořadí slov a flexe 168
8 Obecné problémy 169
8.1 Distribuce 169
8.2 Entropie a velikost inventáře 170
8.3 Aplikace teoretického rozdělení 170
8.4 Vytváření hypotéz pomocí faktorové analýzy 171
8.5 Ikoničnost 172
8.6 Tvoření indexů 173
8.7 Menzerathův zákon 174
8.8 Naranan-Balasubrahmanyanovo rozdělení 175
8.9 Ordovo kritérium 176
8.10 Index opakování a entropie 178

8.11 Velikost výběrového souboru 179
8.12 Problém nekonečna 180
8.13 Těsnost/koheze 181
8.14 Zipfův a Zipfův-Mandelbrotův zákon 183
9 Výzkumné projekty 184
9.1 Frumkinové zákon (výskyt slov v daných pasážích textu) 184
9.2 Skaličkův typologický systém 187
9.3 Synonymie 190
9.4 Frekvence slov a příbuzné vlastnosti 194

Předmluva

Tato kniha je prvním dílem série *Kvantitativní lingvistika. Vybrané problémy*, jež se zabývá výzkumnými záměry, problémy, otázkami, hypotézami a cvičeními z širokého spektra oborů kvantitativní lingvistiky. Jen málo zde prezentovaných problémů bylo již dříve studováno v předchozích výzkumech, každý z nich je však hoden vědeckého zájmu a může vést k poznatkům, které mohou přispět k budování komplexní lingvistické teorie.

Prezentované problémy mají různý stupeň obtížnosti a vyžadují tedy také různě velké úsilí při svém řešení. Mnohé z nich mohou být nápomocny studentům při výběru témat diplomových prací, učitelům při hledání vhodných cvičení pro výuku nebo vědcům při hledání nových výzkumných záměrů. Většina hypotéz umožňuje přinést originální příspěvek do jedné z oblastí kvantitativní lingvistiky nalezením prvních odpovědí na dané otázky, vyřešením problémů, užitím nových metod či přístupů, popř. aplikací již existujících metod na nová data.

Velká většina problémů se zabývá vzájemným vztahem dvou či více lingvistických entit. Čtenář je veden k tomu, aby formuloval přesné definice, způsoby kvantifikace a měření, shromáždil data, provedl testy, našel empirickou funkci nebo odvodil funkci z určitých teoretických předpokladů. Úplné řešení problému nicméně není vždy vyžadováno. Případy, kdy mohou být řešení či metody nalezeny v uvedené literatuře, by čtenáře měly povzbudit k jejich ověřování v jiných jazycích, žánrech atd., popř. k nalezení alternativního řešení.

Jednotlivé problémy jsou v celé knize prezentovány v jednotné formě následujícím způsobem: (1) Ke každé *hypotéze* nebo *problému* jsou uvedeny odkazy na literaturu, která by měla být prostudována. Tyto zdroje často poskytují předběžné analýzy a odkazy na další literaturu. (2) V *postupu* jsou navrženy jednotlivé doporučené kroky analýzy. V některých případech je prezentována i důkladná analýza daného problému.

(3) V *odkazech na literaturu* čtenář může najít první zmínky o dané problematice nebo také její hlubší analýzy. Na základní literaturu, která by měla být nezbytně nastudována před analýzou daného problému, je odkázáno v textu.

Instrukce u jednotlivých problémů vždy neobsahují potřebné vzorce. V těchto případech je čtenář odkázán na doporučenou literaturu nebo učebnice statistiky.

Následující obecná doporučení by měla být nápomocna k úspěšné práci:

(1) Jazykové příklady nemohou být považovány za důkaz daného jevu, struktury, trendu nebo zákona. Jediným vhodným empirickým základem jsou data z kompletních objektů (např. textů) nebo náhodných výběrových souborů.

(2) Korelační analýza není přijatelná jako výsledek, to samé platí pro jednoduchý test rozdílů mezi objekty. Místo toho by měla být nalezena alespoň empirická funkce.

(3) Ačkoliv je angličtina nebo čeština vhodným jazykovým materiálem, doporučujeme obohatit vaše studium o nejméně jeden další jazyk.

(4) Empirická zjištění jsou často předčasně zobecňována. Relevantní závěry by měly být testovány na několika jazycích, žánrech, autorech atd. (v závislosti na druhu konkrétní hypotézy).

(5) Pojmy, kvantifikace a měření musí být explicitně definovány jednoznačným způsobem. Vyhněte se pojmům, které nelze operacionalizovat s dostatečnou přesností.

(6) Vždy se snažte odvození funkce nebo rozdělení, které chcete použít pro vaše data, odůvodnit rozumnými teoretickými předpoklady. Úvahy týkající se proporcionality mohou být často úspěšné, což se ukázalo u mnoha hypotéz v synergetické lingvistice.

(7) Pokud se funkce nebo rozdělení zdá být neadekvátní pro vaše data, překontrolujte tato data (zdroje, zpracování, množství, technické faktory atd.), výpočty, početní postupy a také vaše předpoklady. Změňte nebo opravte cokoliv, co by mohlo být špatně, a zkuste celý postup zopakovat.

(8) Pokud váš matematický model opět selže: v určitých případech ovlivňují vztah hraniční podmínky (ačkoliv víme, že platí gravitační zákon, pozorujeme, že některé objekty, např. ptáci, k zemi nepadají). Najděte takové hraniční podmínky ve vašem případu a zvažte je jako nezávislé proměnné. Přeformulujte vaši hypotézu a začněte znovu.

(9) Žádná hypotéza by neměla být definitivně zamítnuta, nebo definitivně přijata. Její potvrzení (či odmítnutí) nemá nikdy absolutní platnost.

(10) Abyste si ujasnili své myšlenky, vytvořte graf vyjadřující zkoumaný vztah, který bude obsahovat všechny parametry a požadavky (viz způsob zápisu v synergetické lingvistice).

(11) Mějte na paměti, že data jsou do značné míry jen uměle vytvořené konstrukty. Soubor dat transformuje fakta prostřednictvím hypotéz (nebo alespoň prostřednictvím předpokladů či očekávání) do formy tvrzení. Z toho důvodů by měla být nejdříve formulována jasná a věrohodná hypotéza a teprve potom by se měla vytvořit data.

(12) V případě, že je obtížné stanovit, která proměnná je závislá a která nezávislá, snažte se integrovat obě varianty do většího kontrolního cyklu, popř. alespoň testovat obě možnosti.

(13) Jakmile vyřešíte několik problémů, pokuste se je začlenit do kontrolního cyklu. Doplňte chybějící uzly a hrany podle hypotetických předpokladů a následně je zkuste nalézt empiricky.

(14) Nikdy nepřevádějte základní data na procenta, vždy předkládejte absolutní čísla.

(15) Pokud je problém vyřešen, nepovažujte toto řešení za konečné. Podívejte se na problém jako na část většího celku a snažte se jej popsat z této perspektivy.

(16) Jestliže potřebujete nějakou klasifikaci, neprovádějte ji mechanicky pomocí metod, které máte zrovna po ruce. Místo toho se snažte vytvořit teorii a z ní vyvoďte vhodnou klasifikaci.

(17) Nepoužívejte funkce s mnoha parametry (např. mnohočleny), protože budete muset tyto parametry později interpretovat (tj. držte se pravidla Occamovy břitvy).

(18) Pokud jste lingvista, spolupracujte s programátorem a matematikem. Jestliže jste matematik, měl byste výzkum konzultovat se zkušeným lingvistou, protože i sebelepší matematický model je bez možnosti lingvistické interpretace k ničemu.

(19) Zkoušejte aplikovat vyřešené problémy v této knize na nová data (z jiných jazyků), existující teorie tak mohou být potvrzeny, nebo zamítnuty.

(20) Nepovažujte jazykové jednotky za předem dané. Definujte jednotky takovým způsobem, abyste je mohli použít v hypotézách, a to i v případech, kde jejich segmentace může působit poněkud uměle. Mějte na paměti, že tyto jednotky mohou být užitečné z hlediska teorie pro formulaci zákonů (nikoliv gramatických pravidel).

(21) Vždy preferujte funkce a rozdělení s vhodným teoretickým základem před těmi, které sice vykazují lepší shodu modelu s daty, avšak nemají lingvistický základ. Používejte tedy empirické funkce pouze na začátku výzkumu.

(22) Tato kniha sestává z devíti kapitol. Obsah jednotlivých kapitol není striktně homogenní, ale poskytuje relativně široké spektrum problémů, které mohou být řešeny kvantitativními metodami. Problémy jsou v každé kapitole seřazeny abecedně. Některé problémy byly analyzovány detailněji než jiné. Není třeba číst kapitoly a problémy postupně, každý si může vybrat pouze ty části, které odpovídají jeho zájmu a zaměření.

PODĚKOVÁNÍ

Jsme velmi zavázáni Reinhardu Köhlerovi, který pečlivě pročetl celou knihu, zlepšil některé argumentace a dal nám mnoho cenných rad.

1 Fonologie a písmo

1.1 PŘÍZVUK A FREKVENCE

Hypotézy

„…nejfrekventovanější slova jsou zpravidla nepřízvučná." (Zipf 1935: 131).
„…přízvuk má tendenci (1) se vyhýbat slovům s vysokou četností a (2) tíhne ke slovům s méně obvyklým užitím…" (Zipf 1935:132).
„…přízvuk má tendenci se realizovat na morfémech s největším průměrným intervalem (vlnovou délkou), tzn. na morfémech s nejnižší relativní frekvencí…" (Zipf 1935: 136).

Postup

Vezměte text, přečtěte jej nahlas a rozdělte slova na přízvučná a nepřízvučná. Potom zjistěte četnost slov v každé z těchto dvou skupin z korpusu nebo frekvenčního slovníku. V rámci každé skupiny seřaďte slova sestupně podle frekvence a proveďte neparametrický rankový test, který bude zobrazovat, že tyto dvě skupiny slov (přízvučných a nepřízvučných) nenáleží do stejné „přízvukové populace". Zkuste provést tento test v několika jazycích. Pokud se vyskytne slovo, které může být přízvučné i nepřízvučné v různých kontextech, zařaďte jej do obou skupin nebo jej odstraňte z analyzovaného vzorku.

Literatura

Zipf, G. K. (1935). *The psycho-biology of language. An introduction to dynamic philology.* Boston: Houghton Mifflin.

1.2 KANONICKÁ STRUKTURA SLOVA

Hypotéza

Vztah mezi slabičnou a fonetickou délkou kanonických tvarů je lineární.

Postup

Kanonické tvary jsou slova, jejichž fonémy byly redukovány na konsonanty a vokály. Takto získáme formy jako *V, KV, VK, KVK, KVV* atd. Podívejte se do slovníku a přepište všechna slova do formy kanonických tvarů. Pokud používáte počítačový program, dejte si pozor na diftongy a kombinované grafémy (např. E. <sh>, G. <sch>, <ei> atd.). Vytvořte dvoudimenzionální tabulku, v níž bude počet slabik první proměnnou a počet fonémů druhou proměnnou. Dokažte, že relace <*počet slabik, počet fonémů*> je lineární. Nezapomňte, že *KV* a *VK* patří do stejné skupiny (1 slabika, 2 fonémy), zatímco *KVK* a *KVV* patří do skupiny jiné: <1 slabika, 3 fonémy> a <2 slabiky, 3 fonémy>.

Testujete hypotézu (a) bez zohlednění frekvence a (b) se zohledněním frekvence. V obou případech by měl být výsledkem lineární vztah.

Literatura

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie*. Göttingen: Peust & Gutschmidt.

1.3 KONSONANTY A SHLUKY

Hypotéza

Jazyk s bohatým inventářem konsonantů má zároveň mnoho dlouhých konsonantických shluků (Skalička 1964). Se vzrůstajícím inventářem fonémů však (relativní) počet konsonantických shluků klesá.

Postup

Spočítejte velikost inventáře konsonantů a najděte všechny shluky v daném jazyce. Použijte buď korpus, nebo slovník. Proveďte tento postup u deseti různých jazyků a stanovte funkci závislosti. Data vybírejte jak z jazyků s malým inventářem konsonantů, tak z jazyků s inventářem velkým.

Literatura

Skalička, V. (1964). Konsonantenkombination und linguistische Typologie. *Travaux linguistiques de Prague 1*, 111–114.

1.4 DISTRIBUCE KANONICKÝCH TVARŮ

Problém

Kanonické tvary v předchozím problému (při zohlednění frekvence) mají velmi pravidelnou dvoudimenzionální distribuci, jejímiž nezávislými proměnnými jsou délka slabiky a délka fonému. Zkuste odvodit tuto distribuci teoreticky z nějakých přijatelných předpokladů.

Literatura

žádná

1.5 DISTRIBUČNÍ POČET

Problém

Proveďte kompletní fonémicko-distribuční analýzu (viz Altmann, Lehfeldt 1980) v jazyce, který nebyl dosud takto analyzován. Použijte uvedenou literaturu.

Postup

Použijte slovník nebo korpus a nejprve vytvořte všechny různé sekvence dvou fonémů (ne písmen). Proveďte klasický Harary-Paperův výpočet za použití nových indexů. Potom spočítejte frekvence všech sekvencí a proveďte frekvenční fonémicko-distribuční analýzu. Vypočtěte různé indikátory. Zjistěte, zda fonémy s vysokou kolokací mají rovněž vyšší frekvenci. Popište tuto závislost (viz část 4.5, „Kotextualita a frekvence") a postulujte hypotézu.

Literatura

Altmann, G., Lehfeldt, W. (1972). Typologie der phonologischen Distributionsprofile. *Beiträge zur Linguistik und Informationsverarbeitung 22*, 8–32.

Altmann, G., Lehfeldt, W. (1980). *Einführung in die Quantitative Phonologie*. Bochum: Brockmeyer.

Birnbaum, H. (1967). Syntagmatische und paradigmatische Phonologie. In: Hamm, J. (ed.), *Phonologie der Gegenwart*. Graz u. a.: Böhlau, 307–352.

Doležel, L., Průcha, J. (1966). A statistical law of grapheme combinations. *Prague Studies in Mathematical Linguistics 1*, 33–43.

Greenberg, J. H. (1964). Nekotorye obobščenija, kasajuščiesja vozmožnych načal'nych i konečnych posledovatel'nostej soglasnych. *Voprosy jazykoznanija 4*, 41–65.

Harary, F., Paper, H. H. (1957). Toward a general calculus of phonemic distribution. *Language 33*, 143–169.

Hirsch-Wierzbicka, L. (1971). *Funktionelle Belastung und Phonemkombination*. Hamburg: Buske.

Kempgen, S. (1995). Phonemcluster und Phonemdistanzen (im Russischen). *Slavistische Linguistik 1994*, 197–221.

Kempgen, S. (1999). Modellbedingte Distributionsbeschränkungen in der Phonologie. In: Grünberg, K., Potthoff, W. (eds.), *Ars Philologica. Festschrift für Baldur Panzer zum 65. Geburtstag*. Frankfurt am Main. u. a.: Lang, 179–184.

Kempgen, S. (2001). Assoziativität der Phoneme im Russischen. In: Uhlířová, L. et al. (ed.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of L. Hřebíček*. Trier: Wissenschaftlicher Verlag, 124–135.

Lehfeldt, W. (1972). Phonologische Typologie der slavischen Sprachen. *Die Welt der Slaven 17*, 318–340.

Lehfeldt, W. (2005). Phonemdistribution. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin: de Gruyter, 181–190.

Saporta, S. (1955). Frequency of consonant clusters. *Language 31*, 25–31.

Trnka, B. (1936). General laws of phonemic combinations. *Travaux du Cercle Linguistique de Prague 6*, 57–62.

Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. *Travaux du Cercle Linguistique de Prague 7*. Prague. [Nachdruck: Nendeln: Kraus, 1968].

Vogt, H. (1942). The structure of the Norwegian monosyllable. In: *Norsk Tidsskrift for Sprogvidenskap 12*, 5–29.

Vogt, H. (1954). Phoneme classes and phoneme classification. *Word 10*, 28–34.

1.6 DISTRIBUČNÍ MEZERY

Hypotéza

Čím větší je počet fonémů v inventáři, tím menší je podíl možných kombinací fonémů, tzn. tím větší je podíl strukturních mezer.

Postup

Nejdříve vyřešte problém v části 1.5, „Distribuční počet", potom mechanicky spočítejte počet strukturních mezer, tj. spočítejte počet nerealizovaných kombinací fonémů. Tento počet dejte do vztahu k velikosti inventáře fonémů.

Vzhledem k tomu, že data jsou dostupná z předchozího problému, není třeba analyzovat nová data. Formalizujte tento vztah.

Literatura

Schulz, K.-P., Altmann, G. (1988). Lautliche Strukturierung von Spracheinheiten. *Glottometrics 9,* 1–48.

1.7 VÝVOJ KOMPLEXITY PÍSMA

Problém

Grafické znaky dvou historických období se ve vývoji jakéhokoliv písma liší ve své komplexitě. Dokažte, že tato změna není lineární.

Postup

Zvolte dvě historická období jednoho písma, např. bráhmí a dévanágarí, starou a moderní čínštinu, japonské kandži a hiraganu (nebo katakanu), starou a novější asyrštinu, egyptské hieroglyfy a meroitické písmo atd. Změřte komplexitu jednotlivých znaků. Komplexitu starší varianty považujte za proměnnou x a novější za proměnnou y. (a) Dokažte, že vztah není lineární. (b) Zkuste najít vhodnou funkci.

Literatura

Hegenbarth-Reichardt, I., Altmann, G. (2008). On the decrease of complexity from hieroglyphs to hieratic symbols. In: Altmann, G., Fan, F. (eds.), *Analyses of script. Properties of characters and writing systems*. Berlin, New York: de Gruyter, 101–110.

1.8 VYUŽITÍ KANONICKÝCH TVARŮ

Problém

Najděte funkci vyjadřující využití kanonických tvarů.

Postup

Pracujte s fonémickou délkou kanonických tvarů z předchozího problému (tj. marginální distribucí). Použijte pouze jednotlivé typy, nikoliv jejich četnosti. Vzhledem k tomu, že jsou zde jen dva různé prvky (*V, K*), nelze teoreticky získat více než 2 prvky o délce 1, *V* a *K* (připouštíme také varianty *K, KK, KKK* atd., některé z nich existují např. ve slovanských jazycích). Teoreticky jsou možné $2^2 = 4$ varianty prvků o délce 2 (*VV, VK, KV, KK*) a obecně 2^k varianty o délce *k*. Vzhledem k tomu, že počty variant jsou známy z předchozího problému a teoretické počty mohou být odvozeny, vytvořte způsob měření využití daných variant a najděte funkci vyjadřující využití kanonických tvarů. Pokud je to možné, porovnejte tyto funkce v několika jazycích.

Literatura

Altmann, G. (2005). Phonic word structure. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 191–198.

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie*. Göttingen: Peust & Gutschmidt, 48nn.

1.9 FREKVENCE PÍSMEN

Problém

Najděte obecný model pro rankovou frekvenční distribuci písmen.

Postup

Rosenbaum a Fleischmann (2002, 2003) prezentovali mnoho distribucí písmen a diakritických znamének z evropských jazyků.

(1) Pokuste se ukázat, že všechny mají stejné teoretické rozdělení.

(2) Tito autoři představili rovněž pořadí latinských písmen v analyzovaných jazycích. Použijte různé testovací metody pro zjištění podobnosti mezi jazyky na základě četnosti písmen. Pokud budete úspěšní, použijte vhodnější data a rozšiřte tento výzkum.

(3) Z výsledků vyvoďte obecné závěry.

Literatura

Rosenbaum, R., Fleischmann, M. (2002). Character frequency in multilingual corpus 1 – Part 1. *Journal of Quantitative Linguistics 9*(3), 233–260.

Rosenbaum, R., Fleischmann, M. (2003). Character frequency in multilingual corpus 1 – Part 2. *Journal of Quantitative Linguistics 10*(1), 1–39.

1.10 MĚŘENÍ DISTINKCE

Problém

Definujte měření distinkce mezi jednotlivými písmy.

Postup

Vezměte si runové písmo a vypočítejte jeho distinkce aplikací metody Antiće, Altmanna (2005). Vezměte další runové písmo a vzájemně je porovnejte (viz Mačutek 2008). Pokud najdete nějaké rozdíly, popište je. Navrhněte způsob výpočtu distinkce u písma ogham.

Literatura

Antić, G., Altmann, G. (2005). On letter distinctivity. *Glottometrics 4*, 46–53.

Mačutek, J. (2008). Runes: complexity and distinctivity. *Glottometrics 16*, 1–16.

1.11 MĚŘENÍ ORNAMENTALITY

Problém

Ornamentalita není inherentní vlastností písma, je stanovena výhradně naším rozhodnutím při vytváření tohoto pojmu. Nekoresponduje s žádnou reálnou vlastností, ale může být transformována do reálných objektů. Pokuste se najít metodu pro měření ornamentality písma.

Postup

Zde jsou tři možnosti měření:

(1) Vytvořte nějakou stupnici měření a spoléhejte se na úsudek testovacích osob. Tato metoda již byla vyzkoušena.

(2) Měřte ornamentalitu jako redundantní část znaku, která není nezbytná pro jeho identifikaci.

(3) Navrhněte novou objektivní metodu inspirovanou kaligrafií nebo uměním.

Literatura

Altmann, G., Fan, F. (eds.) (2008). *Analyses of script. Properties of characters and writing systems*. Berlin, New York: de Gruyter.

1.12 FREKVENCE FONÉMŮ A SLOV

Hypotéza

„… lexikální jednotky s nízkou frekvencí obsahují častěji fonémy, které mají nízkou frekvenci, než lexikální jednotky s vysokou frekvencí." (Frisch, Large, Zawaydeh, Pisoni 2001: 167).

Postup

Vzhledem k tomu, že frekvence fonémů je přímou funkcí frekvence slov, hypotéza je zřejmá. Pokuste se ji zpřesnit. Spočítejte frekvence fonémů a frekvence slovních tvarů v korpusu. Potom u každého slovního tvaru spočítejte frekvenci jednotlivých fonémů a vypočtěte její průměr. Pokud je hypotéza pravdivá, měli byste získat jednoduchou závislostní funkci. Zkuste vytvořit tuto funkci, tj. závislost průměrné frekvence fonému na frekvenci slovního tvaru. Pokud je to možné, proveďte výpočet na více jazycích a výsledky porovnejte. Pokuste se vyvodit obecné závěry.

Literatura

Frisch, S. A., Large, N. R., Zawaydeh, B., Pisoni, D. B. (2001). Emergent phonotactic generalizations in English and Arabic. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 159–179.

Frauenfelder, U. H., Baayen, R. H., Hellwig, F. M., Schreuder, R. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language 32*, 781–804.

Landauer, T. K., Streeter, L. A. (1973). Structural differences between common and rare words. Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior 12*, 119–131.

1.13 INVENTÁŘ FONÉMŮ A DÉLKA SLOVA

Hypotéza

Čím větší je inventář fonémů, tím menší je průměrná délka slova (Nettle 1995).

Postup

Nettle (1995) vypočítal tento vztah v 10 jazycích, přičemž měřil délku slova ve fonémech.

(1) Porovnejte více jazyků, aby mohla být hypotéza potvrzena nebo modifikována.

(2) V případě, že se hypotéza ukáže být nedostačující, přidejte další vlastnosti jazyka a vytvořte funkci se dvěma nezávislými proměnnými. Další vlastností může být např. rozsah distribuce fonémů (asociativnost fonémů, počet fonémických bigramů v jazyce).

(3) Testujte hypotézu na textech (nikoliv slovníku).

(4) Použijte průměrnou délku slabik ve slově jako závislou proměnnou a ujistěte se o platnosti hypotézy.

(5) Určete vlastnosti, které by mohly mít vliv na délku slova v jazyce.

(6) Testujte hypotézu za použití průměrné délky morfu jako závislé proměnné.

Literatura

Hockett, C. F. (1958). *A course in modern linguistics*. Toronto: McMillan.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

Nettle, D. (1995). Segmental inventory size, word length, and communicative efficiency. *Linguistics 33*, 359–367.

Weber, S. (2005). Zusammenhänge. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 214–226.

1.14 MOCNINNÝ ZÁKON

Problém

Zkuste aplikovat Narananem a Balasubrahmanyanem modifikovaný mocninný zákon a modifikované mocninné rozdělení částečných součtů na jazyková data.

Postup

Získejte co nejvíce rankových frekvenčních distribucí fonémů/písmen. Aplikujte tyto distribuce na vaše data. Pokud není dostupný vhodný software, pokuste se odvodit metodu pro odhad parametrů za použití frekvencí nejnižších ranků. Zkuste interpretovat distribuci částečných součtů v lingvistických termínech.

Literatura

Naranan, S., Balasubrahmanyan, V. K. (2005). Power laws in statistical linguistics and related systems. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 716–738.

1.15 POŘADÍ TYPŮ SLABIK

Problém

Zjistěte rankovou frekvenční distribuci slabik.

Postup

Schiller et al. (1996) představili procentní zastoupení kanonických tvarů slabik v holandštině, a to jak v typech, tak v tokenech, a seřadili je sestupně podle frekvence. Pokuste se najít vztah mezi pořadím a procentním zastoupením. Použijte funkci, nikoliv distribuci. Vyvoďte z výsledků závěry.

Literatura

Schiller, N. O., Meyer, A. S., Baayen, R. H., Levelt, W. J. M. (1996). A comparison of lexeme and speech syllables in Dutch. *Journal of Quantitative Linguistics 3(1)*, 8–28.

1.16 KOMPLEXITA PÍSMA

Problém

Existuje několik možností, jak měřit komplexitu písma (znaku): průsečíková metoda, škálovací metoda, Bézierovy křivky, počítání tahů, počítání pixelů, fraktální dimenze atd. Pokuste se definovat nový způsob měření nebo použijte všechna existující měření pro jedno písmo a porovnejte je.

Postup

Vzhledem k tomu, že jednotlivá měření pokrývají jen izolované vlastnosti písma, pokuste se vytvořit takový způsob měření, který bude brát v potaz (a) tvar čar, (b) délku čar, (c) směr čar, (d) spojení čar. Aplikujte měření ve fontu Arial a porovnejte je s dosavadními výsledky.

Vezměte si maďarské runy, které je možné najít na internetu, a vypočítejte komplexitu každého znaku. Použijte známé metody měření komplexity znaku. Potom vezměte další runová písma a porovnejte je s maďarským. Zejména zvažte písmo ogham.

Literatura

Altmann, G. (2004). Script complexity. *Glottometrics 8*, 68–73.

Altman, G., Fan, F. (eds.) (2008). *Analyses of script. Properties of characters and writing systems.* Berlin, New York: de Gruyter.

Mačutek, J. (2008). Runes: complexity and distinctivity. *Glottometrics 16*, 1–16.

1.17 ZJEDNODUŠOVÁNÍ PÍSMA

Problém

Prokažte, že ve vývoji písma docházelo k postupnému zjednodušování a že byl tento proces lineární.

Postup

(1) Použijte měření komplexity písma navržené Altmannem (2004). Použijte Haarmannovu knihu (1990) nebo Omniglot (internet) a vyberte tabulku, ve které je prezentováno určité období písma (např. aramejské písmo, s. 301). Vypočítejte komplexitu „staré" a „nové" podoby a míru zjednodušení.

(2) Proveďte stejný postup u japonských písem kandži a dále hiragana a katakana, která se z nich vyvinula.

(3) Vypočtěte proces změn komplexity znaku u nejstarších a moderních čínských znaků, výsledky porovnejte.

(4) Zjistěte změny komplexity ve vývoji asyrštiny od nejstaršího klínového písma až po nejnovější formy.

(5) Vyberte si několik run (nalézt je můžete na internetu, např. Omniglot) a vypočítejte jejich komplexitu. Zjistěte, zda má jejich průměrná komplexita statisticky rovnocenná a najděte příčinu tohoto faktu. Pokud se statisticky liší, zjistěte, zda jejich stáří (doba prvního výskytu) ovlivňuje komplexitu. Pokuste se nalézt kauzální, psychologické, sociální a další faktory, které způsobují rozdíly v komplexitě. Vypočítejte, zda je zjednodušování znaků lineární, nebo zde existuje jiný trend.

(6) Porovnejte hieroglyfy s meroitským písmem.

Literatura

Altmann, G. (2004). Script complexity. *Glottometrics 8*, 68–73.

Haarmann, H. (1990). *Universalgeschichte der Schrift*. Frankfurt am Main: Campus.

Hegenbarth-Reichardt, I., Altmann, G. (2008). On the decrease of complexity from hieroglyphs to hieratic symbols. In: Altmann, G., Fan, F. (eds.), *Analyses of script. Properties of characters and writing systems*. Berlin: de Gruyter, 105–114.

Mačutek, J. (2008). Runes: complexity and distinctivity. *Glottometrics 16*, 1–16.

1.18 FREKVENCE SLABIK

Hypotéza

Ranková frekvenční distribuce slabik má stejné vlastnosti jako ranková frekvenční distribuce slov.

Postup

Použijte data z korpusu. (a) Pokud je to možné, proveďte fonologickou transkripci korpusu nebo (b) použijte jeho psanou podobu. V obou případech rozdělte slova na slabiky a vypočítejte frekvenční distribuci jednotlivých slabik (nikoliv jejich kanonických tvarů!). Jestliže máte k dispozici software na segmentaci slov na slabiky, použijte jej. Vytvořte rankovou frekvenční distribuci slabik a zkuste ji porovnat s distribucí slov. Pokud získáte rozdílné výsledky, hledejte jejich příčinu. Zkuste změnit způsob segmentace, stanovte hraniční podmínky a začleňte je do teoretického rozdělení, pokuste se odvodit teoretické rozdělení na základě kombinatorických předpokladů.

Literatura

Bektaev, K. B. (1973), Alfavitno-častotnyj slovar' slogov kazachskogo jazyka. In: *Statistika kazachskogo teksta I. Trudy gruppy „Statistiko-lingvističeskoe issledovanie i avtomatizacija" III.* Alma-Ata: Nauka, 566–611.

Schiller, N. O., Meyer, A. S., Bayen, R. H., Levelt, W. J. M. (1996). A comparison of lexeme and speech syllables in Dutch. *Journal of Quantitative Linguistics* 3(*1*), 8–28.

1.19 STRUKTURA SLABIK

Problém

Popis struktury slabik zahrnuje několik problémů, které je třeba řešit postupně.

Postup

(1) Vytvořte inventář slabik jazyka (srov. část 1.18, „Frekvence slabik").

(2) Vyřešte problém v části 1.18, „Frekvence slabik“, za použití korpusu.

(3) Prozkoumejte vztah mezi frekvencí slabik a jejich délkou. Vzhledem k tomu, že slabiky jsou poměrně krátké, bude snadné najít vhodnou funkci.

(4) Slabiky obsahují počáteční a koncovou jednotku (onset a kodu). Prozkoumejte jejich symetrii a asymetrii, vytvořte indikátor symetrie. Zjistěte vlastnosti tohoto indikátoru.

(5) Porovnejte inventář slabik s inventářem fonémů (v několika jazycích). Je zde nějaká závislost? Pokud ano, jaká?

(6) Pokuste se stanovit pravidlo využití, tj. spočítejte počet všech potenciálních slabik o délce x a spočítejte počet skutečně realizovaných slabik. Vytvořte způsob měření využití inventáře slabik.

(7) Zjistěte, zda měření využití v (6) nemá souvislost s fonologickým typem jazyka.

(8) Vytvořte fonologická pravidla pro tvoření počátku a konce slabiky (onsetu a kody).

(9) Testujte existenci konsonantické harmonie mezi počátky a konci slabik (onsety a kodami).

Literatura

Berg, T. (1994). The sensitivity of phonological rimes to phonetic length. *Arbeiten aus Anglistik und Amerikanistik 19*, 63–81.

Booij, G. (1995). *The phonology of Dutch*. Oxford: Clarendon.

Bortoloni, U. (1976). Tipologia sillabica d´italiano. Studio statistico. In: Simone, R., Vignuzzi, U., Ruggiero, G. (eds.), *Studi di fonetica e fonologia. Atti del convegno internazionale di studi. Padova 1 e 2 ottobre 1973*. Roma, 5–22. [Pubblicazioni della Società di Linguistica Italiana 9].

Browman, C. P., Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica 45*, 140–155.

Delattre, P. (1966). A comparison of syllable length conditioning among languages. *International Review of Applied Linguistics 4,* 183–198.

Derwing, B. L., Yoon, Y. B., Cho, S. W. (1993). The organization of the Korean syllable: Experimental evidence. In: Clancy, O. M. (ed.), *Japanese/ Korean Linguistics, Vol. 2.* Stanford: Center for the Study of Language and Information, 223–238.

Derwing, B. L., Dow, M. L., Nearey, T. M. (1988). Experimenting with syllable structure. In: Powers, J., de Jong, K. (eds.), *Proceedings of the Fifth Eastern States Conference on Linguistics.* Columbus: Ohio State University, 83–94.

Eisenberg, P., Ramers, K.-H., Vater, H. (eds.) (1992). *Silbenphonologie des Deutschen.* Tübingen: Narr.

Fallows, D. (1981). Experimental evidence for English syllabification and syllable structure. *Journal of Linguistics 17,* 309–317.

Fowler, C. A., Treiman, R., Gross, J. (1993). The structure of English syllables and polysyllables. *Journal of Memory and Language 32*, 115–140.

Goldinger, S. D., Luce, P. A., Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language 28*, 501–518.

Grainger, J. (1992). Orthographic neighbourhoods and visual word recognition. In: Frost, R., Katz, L. (eds.), *Orthography, phonology, morphology and meaning*. Amsterdam: Elsevier, 131–146.

Hall, T. (1962). *Syllable structure and syllable related processes in German.* Tübingen: Narr.

Pike, K., Pike, E. (1947). Immediate constituents of Mazateco syllables. *International Journal of American Linguistics 13*, 78–91.

Portele, T. (1995). The influence of the syllable boundary on consonant--consonant realizations. *Proceedings of the International Congress of Phonetic Sciences. Vol 2.* Stockholm, 594–597.

Pulgram, E. (1970). *Syllable, word, nexus, cursus.* The Hague: de Gruyter.

Schiller, N. O., Meyer, A. S., Levelt, W. J. M. (1997). The syllabic structure of spoken words: evidence from the syllabification of intervocalic consonants. *Language and Speech 40*(2), 103–140.

Selkirk, E. O. (1982). The syllable. In: Hulst, H. van der, Smith, N. (eds.), *The structure of phonological representations. Part II*. Dordrecht: Foris, 337–383.

Sommer, B. (1970). An Australian language without CV syllables. *International Journal of American Linguistics 36*, 57–58.

Treiman, R., Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory and Language 27*, 87–104.

Treiman, R., Fowler, C. A., Gross, J., Berch, D., Weatherston, S. (1995). Syllable structure or word structure? Evidence for onset and rime units with disyllabic and trisyllabic stimuli. *Journal of Memory and Language 34*, 132–155.

Treiman, R., Zukowski, A. (1990). Toward an understanding of English syllabification. *Journal of Memory and Language 29*, 66–85.

Vennemann, T. (1988). *Preference laws for syllable structure and the explanation of sound change*. Berlin: de Gruyter.

Vennemann, T. (1972). Zur Silbenstruktur der deutschen Standardsprache. In: Vennemann, T. (ed.), *Silben, Segmente, Akzente*. Tübingen: Niemeyer.

1.20 TENDENCE K VOKÁLOVÉ HARMONII

Hypotéza

V dvouslabičných morfémech existuje tendence k vokálové harmonii.

Postup

Vezměte dvojslabičná slova ze slovníku. Pro některé jazyky výše zmíněná hypotéza platí (např. indonéské jazyky). Liší se od obvyklé deterministické vokálové harmonie samohlásek v afixech, např. v maďarštině. Zkuste zjistit, zda platí ve vámi zkoumaném jazyce.

Jsou dvě možnosti: (a) vokál v první slabice se signifikantně kombinuje se stejným vokálem v druhé slabice. (b) Některé vokály se signifikantně kombinují jen s určitými vokály a naopak se vyhýbají kombinaci s jinými. Použijte vhodný test pro zjištění existence „tendence k harmonii".

Literatura

Altmann, G. (1987). Tendenzielle Vokalharmonie. *Glottometrika 8*, 104–112.

Schulz, K. P., Altmann, G. (1988). Lautliche Strukturierung von Spracheinheiten. *Glottometrika 9*, 1–48.

1.21 DVOJDIMENZIONÁLNÍ STRUKTURA SLABIK

Problém

Pokuste se najít dvojdimenzionální strukturu slabik v evropských jazycích.

Postup

Nejdříve připravte seznam všech možných slabik v daném jazyce. Spočítejte kanonické typy, tj. počet druhů *V, VK, KV, KKV,...* a zaneste jejich četnosti do tabulky, kde první sloupec obsahuje konsonanty před vokály a první řádek konsonanty za vokály, a to následujícím způsobem:

TABULKA 1.21.1
Frekvence kanonických typů

	V	VK	VKK	VKKK	...
V					
KV					
KKV					
KKKV					
...					

Křížení *KV* a *VK* znamená slabiku typu *KVK*. Testujte hypotézu s vašimi daty v distribuci

$$P_{ij} = \frac{a^i b^j}{(i!)^k (j!)^m} P_{00}, \;\; i, j = 0, 1, \ldots ,$$

kde P_{ij} je pravděpodobnost slabiky v řadě *i* a sloupci *j*; *a, b, k, m* jsou parametry a P_{00} je pravděpodobnost slabik typu *V*. Postup můžete nalézt v Zörnig, Altmann (1993).

Pokud najdete odchylku od tohoto modelu, modifikujte model vhodným způsobem nebo vytvořte nový model založený na jiných předpokladech.

Pokuste se analyzovat několik jazyků a nalezněte vedlejší fonologické vlastnosti, které ovlivňují velikost parametrů.

Literatura

Lee, S.-O. (1986). An explanation of syllable structure change. *Korean Language Research* 22, 195–213.

Vennemann, T. (ed.) (1982). Zur Silbenstruktur der deutschen Standardsprache. *Silben, Segmente, Akzente*. Tübingen: Narr, 261–305.

Zörnig, P., Altmann, G. (1993). A model for the distribution of syllable types. *Glottometrika 14*, 190–196.

1.22 DÉLKA SLOVA A SUPRASEGMENTÁLNÍ JEDNOTKY

Hypotéza

Čím více suprasegmentálních jednotek daný jazyk má, tím kratší je průměrná délka slova (Kempgen 1990: 119).

Postup

Prozkoumejte více různých jazyků, které mají suprasegmentální jednotky (různé tóny, přízvuky, různou délku vokálů) pro rozlišení slov. Vypočítejte průměrné délky slov a najděte výše uvedenou závislost. Porovnejte analyzované jazyky s těmi, které nemají suprasegmentální jednotky.

Literatura

Kempgen, S. (1990). Akzent und Wortlänge: Überlegungen zu einem typologischen Zusammenhang. *Linguistische Berichte 126*, 115–134.

2 Gramatika

2.1 BEHAGELŮV ZÁKON

Hypotéza

V korpusu platí, že čím větší je rozdíl x mezi délkou dvou po sobě následujících předložkových frází, tím větší je pravděpodobnost $p(x)$, že kratší předložková fráze předchází delší (Hoffmann 1999: 113). Srov. problém v části 5.8, „Frekvence a pořadí ve frázích".

Postup

Tato hypotéza se zdá být v rozporu s Fenk-Oczlonovou hypotézou, ale nemusí tomu tak nezbytně být. Nejdříve vymezte pojem „předložková fráze", potom zkuste vyjádřit formálně hypotézu. Vyřešte nezbytné záležitosti a použijte korpus jako zdroj dat. Neomezujte se na němčinu nebo angličtinu – data z těchto jazyků jsou jednoduše dostupná – raději získejte data z jiných jazyků. Viz literaturu.

Literatura

Allen, K. (1987). Hierarchies and the choice of left conjuncts (with particular attention to English). *Journal of Linguistics 23*, 51–71.

Bock, J. K., Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition 21*, 47–67.

Cooper, W. E., Ross, J. R. (1975). Word order. In: Grossman, R. E., San, L. J., Vance, T. J. et al. (eds.), *Papers from the parasession on functionalism*. Chicago: Chicago Linguistic Society, 63–111.

Edmondson, J. A. (1985). Biological foundation of language universals. In: Bailey, C. J., Harris, R. (eds.), *Developmental mechanisms of language*. Oxford: Pergamon, 109–130.

Ertel, S. (1977). Where do the subjects of sentences come from? In: Hillsdale, N. J., *Sentence production: developments in research and theory*. Erlbaum, 141–186.

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics 27,* 517–556.

Fenk-Oczlon, G. (1983). Ist die SVO-Wortfolge die ‚natürlichste'? *Papiere zur Linguistik 29,* 23–32.

Fenk-Oczlon, G. (1987). *Frequenz und Wortfolge. Am Beispiel von ‚freezes'*. Paper presented at the XIVth International Congress of Linguists.

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 431–448.

Hawkins, J. A. (1983). *Word order universals*. San Diego: Academic Press.

Hawkins, J. A. (1990). A parsing theory of word order universals. *Linguistic Inquiry 21(2),* 223–261.

Hawkins, J. A. (1992). Syntactic weight versus information structure in word order variation. In: Jacobs, J. (ed.), *Informationsstruktur und Grammatik*. Opladen: Westdeutscher Verlag.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge: University Press.

Hoffmann, Ch. (1999). Word order and the principle of „Early Immediate Constituents" (EIC). *Journal of Quantitative Linguistics 6,* 108–116.

Kelly, M. H., Bock, K. J., Keil, F. C. (1986). Prototypicality in a linguistic context: effects on sentence structure. *Journal of Memory and Language 25,* 59–74.

Kuno, S. (1979). On the interaction between syntactic rules and discourse principles. In: Bedell, G., Kobayashi, E., Muraki, M. (eds.), *Explorations in linguistics: Papers in honor of Kazuko Inoue*. Tokyo: Kenkyusha, 279–304.

Malkiel, Y. (1959). Studies in irreversible binomials. *Lingua,* 113–160.

Mayerthaler, W. (1981). *Morphologische Natürlichkeit*. Wiesbaden: Akademische Verlagsgesellschaft Athenaion.

Pinker, S., Birdsong, D. (1979). Speakers´ sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior 18*, 497–508.

Ross, J. R. (1980). Ikonismus in der Phraseologie. *Zeitschrift für Semiotik 2*, 39–56.

2.2 SPOLUVÝSKYT A KOHEZE

Hypotéza

„…syntaktická koheze je přímým důsledkem frekvence spoluvýskytu: slova, která jsou častěji použita společně, inklinují k tomu se spojit a také se u nich projevuje silnější tendence k hláskovým změnám, které probíhají při spojování slov (liaison)." (Bybee 2001: 338; srov. také s. 343).

Postup

Nejdříve definujte přesnou metodu měření stupňů koheze (viz také část 2.5, „Frekvence a koheze"). Potom spočítejte spoluvýskyty slov v korpusu. Porovnejte počet spoluvýskytů se stupněm koheze. Pokud hypotéza neplatí, hledejte hraniční podmínky, za kterých by platit mohla.

Literatura

Bybee, J. (2001). Frequency effects on French liaison. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 337–359.

2.3 KOTEXTUALITA A VARIACE

Hypotéza

„…pokud se gramatické morfémy vyskytují v různých konstrukcích, vzdalují se jeden od druhého ve fonologické formě, významu a v distribučních vlastnostech." (Bybee 2001: 346n).

Postup

Hypotéza říká, že bohatá kotextualita (bohatá distribuce) způsobuje nárůst počtu různých alomorfů. Vyberte 100 morfémů (autosémantických i synsémantických) a zjistěte jejich kotextualitu v korpusu. Stanovte přímou závislost mezi počtem kontextů a počtem variant. Pokud hypotéza neplatí ve všech případech, zjistěte hraniční podmínky, pokuste se je kvantifikovat a vyjádřete závislost *Formy variant = f(počet kontextů, stupeň další vlastnosti*). Testujte obě hypotézy. Pokud neplatí ani jedna, definujte třetí nezávislou proměnnou.

Literatura

Bybee, J. (2001). Frequency effects on French liaison. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 337–359.

2.4 FREKVENCE A PÁD

Hypotéza

„Čím četnější je pád v určitém jazyce, tím více směřuje k realizaci nulovou formou." (Fenk-Oczlon 2001: 441).

Postup

Vezměte v úvahu jazyk vyjadřující pády. Nerozlišujte pouze opozici nulového a nenulového znaku, ale pokuste se vytvořit metodu, pomocí níž by byl rozsah kódování vyjádřen škálovitě. Latina označuje pády s následujícími koncovkami: *nulová koncovka, -a, -ae, -bus, -ibus, -itis* atd. Určete počet všech substantiv ve všech pádech v korpusu. Pokuste se formálně vyjádřit relaci <*frekvence, rozsah kódování*> pomocí průměrných frekvencí (nebo relativních frekvencí) v rámci každé třídy rozsahu kódování. Pokud to nepůjde, prezentujte své výsledky a vysvětlete důvody. Zkoumejte jeden silně flexivní jazyk a jeden silně aglutinační. Sledujte rozdíly a zkuste je vysvětlit.

Literatura

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistics form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 431–448.

2.5 FREKVENCE A KOHEZE

Hypotéza

„… časté užívání fráze ve standardním úzu činí frázi koherentní, a ta se tak stává samostatnou jednotkou." (Boyland 2001: 395).

Postup

Jelikož ve výše uvedeném smyslu není koheze měřitelná, nejdříve navrhněte jasnou definici pojmu koheze a učiňte ji měřitelnou. Potom shromážděte nejméně 100 frází z korpusu, zjistěte jejich frekvence a tyto frekvence dejte do vztahu k jejich kohezi. Je třeba zmínit, že koheze může být definována různými způsoby. Pokud hypotéza pro vaše data neplatí, zkuste proto

nejdříve změnit definici a způsob měření koheze. Použijte pokud možno korpusová data z jiných jazyků než z angličtiny.

Literatura

Boyland, J. T. (2001). Hypercorrect pronoun case in English? Cognitive processes that account for pronoun usage. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 383–404.

2.6 FREKVENCE A DERIVACE

Hypotéza

„Odvozené slovo má nižší četnost výskytu než jeho základové slovo." (Nagórko-Kufel 1984).

Postup

Na základě frekvenčního slovníku nebo korpusu určete frekvence lemmat. Vezměte třeba 1 000 substantiv a slova z nich odvozená, jež se vyskytly ve slovníku nebo korpusu. Spočítejte frekvence jednotlivých substantiv a slov z nich odvozených. Vytvořte tabulku, ve které je náhodnou proměnnou „rozdíl mezi frekvencí substantiva a frekvencí slova z něj odvozeného", tj. $X = f_{substantivum} - f_{slovo\ odvozené}$. Tato veličina může mít také záporné hodnoty (pokud se odvozené slovo vyskytuje častěji než základní tvar).

Pokuste se najít teoretické rozdělení tohoto rozdílu. Ukažte, že se nejedná o normální rozdělení (testujte např. koeficient šikmosti). Použijte Johnsonovy S_U transformace. Najděte distribuci proměnné $X = |f_{substantivum} - f_{slovo\ odvozené}|$. Zkuste vysvětlit formu distribuce. Pokuste se nalézt diskrétní rozdělení.

Pokračujte analýzou sloves a adjektiv a snažte se podat obecnou teorii. Podívejte se na teorii komplexnosti a teorii příznakovosti. Porovnejte vaše výsledky s výsledky z jiných jazyků.

Literatura

Ginzburg, E. L. (1975). Ob odnom kriterii napravlenija derivacii. *Aktual´nye problemy russkogo slovoobrazovanija* (*Taškent*), 372–376.

Guiraud, P. (1960). *Problèmes et methods de la statistique linguistique*. Paris: PUF.

Harwood, F. W., Wright, A. M. (1956). Statistical study of English word formation. *Language 32*, 260–273.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika 36*, 149–176.

Johnson, N. L., Kotz, S. (1970). *Continuous univariate distributions – Vol. 1*. Boston: Houghton Mifflin.

Nagórko-Kufel, A. (1984). Die Anwendung des Häufigkeitskriteriums bei der Wortbildung. *Glottometrika 6*, 48–64.

2.7 FREKVENCE A NEPRAVIDELNOST

Hypotéza

„… existuje vztah mezi vysokou frekvencí a nepravidelností." (Corbett, Hippisley, Brown, Marriot 2001: 202).

„Čím je nějaká konstrukce četnější, tím větší je pravděpodobnost, že její forma bude zachována, než aby byla nahrazena nějakou produktivnější konstrukcí." (Bybee 2001: 348).

„… to, co je častější…je nepravidelnější." (Corbett, Hippisley, Brown, Marriot 2001: 202).

Postup

Autoři zvažují nepravidelnosti v deklinaci a navrhují způsob, jak nepravidelnost škálovat. (a) Pokuste se transformovat tento problém na případ konjugace nebo jiné gramatické kategorie v jakémkoliv jazyce. (b) Zkuste zobecnit tento problém navržením obecné metody pro škálování odchylek od očekávané hodnoty. (c) Použijte frekvenční slovník slovních tvarů (seřazených podle pořadí), vyberte každé desáté slovo, určete jeho frekvenci a změřte jeho nepravidelnost. Potom zkuste nalézt funkci vyjadřující relaci <*pořadí*, *nepravidelnost*> a analyzujte jej. Přečtěte si diskuzi v citovaném článku a pokuste se zobecnit koncept nepravidelnosti v jazyce.

Vytvořte frekvenční seznam jednotlivých slovesných tvarů z dlouhého textu nebo korpusu. Označte pravidelná slovesa symbolem *R*, nepravidelná (počítá se jakákoliv nepravidelnost, bez škálování) symbolem *I*. Proveďte Wilcoxonův U-test, abyste zjistili, zda platí druhá hypotéza. Potom proveďte to samé se substantivy. Zvolte jazyk se silnou deklinací a potom zkuste problematiku zobecnit.

Literatura

Corbett, G., Hippisley, A., Brown, D., Marriott, P. (2001). Frequency, regularity and the paradigm: A perspective from Russian on a complex relation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 201–226.

Bybee, J. (2001). Frequency effects on French liaison. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 337–359.

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 431–448.

2.8 FREKVENCE A VALENCE

Hypotéza

„…s rostoucí frekvencí slovesa klesá pravděpodobnost pevně stanoveného počtu jeho *argumentových struktur.*"(Thompson, Hopper 2001: 49).
„… čím frekventovanější je sloveso, tím méně je predikovatelný počet jeho argumentů, sloveso s nízkou frekvencí, jako je *to elapse*, má jediný argument, zatímco frekventované sloveso, jako je *to get*, se objevuje s jedním, dvěma nebo třemi tradičně definovanými argumenty…" (Bybee, Hopper 2001: 5).

Hypotéza může být rozšířena: frekventovaná slovesa mají mnoho předložkových (a postpozičních) frází (*get up*, *get in*, *get away*,...).

Postup

Pokuste se zpřesnit hypotézu: *počet argumentů = f(frekvence)*, teoreticky ji odvoďte z přijatelných předpokladů a testujte na 100 (anglických) slovesech, která mají různou frekvenci. Srovnejte výsledek s českými valenčními a frekvenčními slovníky. Zkuste vytvořit závislostní funkci.

Literatura

Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 1–24.

Thompson, S. A., Hopper, P. J. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 49–60.

2.9 FREKVENCE VĚTNÝCH VZORCŮ

Hypotéza

Ranková frekvenční distribuce (rank-frekvence) větných vzorců podléhá Zipf-Mandelbrotovu zákonu (Köhler 2005).

Postup

Abyste mohli testovat hypotézu, analyzujte všechny věty dlouhého textu a přisuďte jim určitou obecnou strukturu v závislosti na typu gramatiky. Potom spočítejte počet vět každého typu v textu a vytvořte jejich rankovou frekvenční distribuci. Testujte, zda tato distribuce odpovídá Zipf-Mandelbrotově rozdělení. Pokud ne, najděte vhodnější typ rozdělení.

Proveďte takové analýzy za použití různých gramatik a okomentujte výsledky. Můžete vyvodit závěr, že nejlepší gramatika je ta, která nejpřesněji odpovídá Zipf-Mandelbrotovu zákonu?

Srov. kapitolu 4, „Textologie", a zkuste převzít některé indikátory, které by mohly – mutatis mutandis – vyjádřit některé syntaktické vlastnosti.

Literatura

Köhler, R. (2005). Quantitative Untersuchungen zur Valenz deutscher Verben. *Glottometrics* 9, 13–20.

2.10 GRAMATIKALIZACE

Problém

Navrhněte metodu pro měření „gramatikalizačního poklesu" (a) od idiomu ke gramatickým pravidlům, (b) od lexémů k flexivním afixům, (c) od frazémů přes kompozita k splynutí (blending) (Hopper, Closs, Traugott 2003).

Postup

Navrhněte stupně (nebo třídy) nezávislosti nebo koheze a pokuste se přiřadit vaše jednotky k jednotlivým třídám nezávislosti či koheze. Vytvořte velký výběrový soubor z korpusu a pokuste se najít nějaké pravidelnosti nebo závislosti.

Literatura

Hopper, P. J., Closs Traugott, E. (2003). *Grammaticalization,* 2nd ed. Cambridge: Cambridge University Press.

2.11 FREKVENCE MORFÉMŮ

Problém

Ranková frekvenční distribuce morfémů se neliší od rankové frekvenční distribuce slov.

Postup

Rozdělte daný text na morfy a vytvořte jejich rankovou frekvenční distribuci. Testujte, zda jsou obvyklá rozdělení vhodná. Srov. problémy „Frekvence slov 1, 2, 3" v kapitole 4.

Literatura

Best, K. H. (2005). Morphlängen. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 255–260.

2.12 POLYSÉMIE MORFÉMU A JEHO FREKVENCE

Krott (1999) prezentovala závislost frekvence morfémů na různých typech jejich polysémie. Ve dvou případech může být tato závislost modelována uspokojivě (u substantiv a sloves), ostatní slovní druhy vykazují velké odchylky. Okomentujte tento jev.

Postup

Je zřejmé, že samotná polysémie nevysvětluje dostatečně značnou část odchylek. Pravděpodobně musí být přidána další nezávislá proměnná, která může být odlišná pro každý jednotlivý slovní druh. Nejdříve se pokuste najít řešení pro každý jednotlivý případ teoreticky na základě nějakého teoretického předpokladu, potom analyzujte dostatečně rozsáhlý výběrový soubor z velkého slovníku a použijte frekvence z korpusu. Pokud to bude nezbytné, přidejte další nezávislé proměnné.

Literatura

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Krott, A. (1999). Influence of morpheme polysemy on morpheme frequency. *Journal of Quantitative Linguistics 6(1),* 58–65.

2.13 MORFOLOGICKÁ PRODUKTIVITA KMENŮ

Problém

Produktivita kmenů (= tvoření odvozenin a kompozit) ve slovníku odpovídá pravidelnému rozdělení pravěpodobnosti.

Postup

Použijte slovník, který obsahuje odvozená a složená slova. Spočítejte počet kmenů, z nichž je utvořeno $x = 0, 1, 2, \ldots$ odvozenin/kompozit a vytvořte empirickou distribuci. Teoretické rozdělení může být vytvořeno aplikací procesu „množení a zániku" (birth-and-death process) (srov. Wimmer, Altmann 1995), přičemž míra množení a zániku nemusí být stejná pro všechny jazyky. Předpokládejte různé míry množení a zániku, vyřešte tento proces a pokuste se celý tento problém zobecnit. Najděte další podmínky, které umožní vysvětlit výběr míry množení a zániku. Vytvořte teorii.

Literatura

Wimmer, G., Altmann, G. (1995). A model of morphological productivity. *Journal of Quantitative Linguistics 2(3)*, 212–216.

2.14 SEKVENČNÍ FREKVENCE SLOVNÍCH TŘÍD

Hypotéza

Kumulativní sekvenční frekvence hlavních slovních tříd (substantiva, slovesa) je konvexní, u pomocných slov (auxiliár) je konkávní.

Postup

Tato hypotéza nebyla dosud testována. Je velmi obecná a bude modifikována mnoha hraničními podmínkami. Přesto by mohla být provedena pilotní studie.

Spočítejte, kolik substantiv se objeví do pozice x ($x = 1, 2, 3, \ldots N$) ve vámi zvoleném textu. Zjistěte kumulativní poziční frekvence substantiv. Substantiva ve výše uvedené hypotéze znázorňují následující sekvenci.

Pozice x	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Substantiva kumulativní	0 0 1 1 1 2 3 4 4 4 4 4 5 6 6 6

Proveďte výpočet pro všechny třídy slov. Potom vezměte jednotlivé řady a aplikujte na ně mocninnou funkci

$$y = ax^b.$$

Pokud je $b > 1$, křivka je konvexní. Pokud je $b < 1$, křivka je konkávní. Jestliže $b = 1$, výsledkem je přímka.

Pokuste se charakterizovat texty, žánry a jazyky vytvořením spektra sekvenčních frekvencí slovních tříd. Zkoumejte slovní druhy v jejich historickém vývoji. Použijte parametry b jednotlivých slovních druhů jako prvky vektoru jednotlivých textů. Vezměte průměrné hodnoty b pro jednotlivé jazyky a porovnejte jejich vektory. Definujte slovní třídy v jednotlivých jazycích porovnatelným způsobem.

Literatura

Ziegler, A., Best, K.-H., Altmann, G. (2001). A contribution to text spectra. *Glottometrics 1*, 97–108.

2.15 KLASIFIKACE SLOVES

Problém

Je adekvátnost aplikace rozdělení pravděpodobnosti na rankové řady kritériem adekvátní klasifikace?

Postup

V kvantitativní lingvistice se často adekvátnost aplikace teoretického rozdělení na ranková data považuje za znak „správnosti“ klasifikace daných jednotek. Použijte data z Levickij, Kiiko a Spolnicka (1996),

kteří klasifikovali německá slovesa do 22 tříd a prezentovali počet sloves v každé třídě.

Seřaďte tuto klasifikaci podle počtu sloves. Potom (a) se pokuste najít teoretické rankové frekvenční rozdělení. Pokud se to nepodaří, (b) zkuste najít induktivně vhodné rozdělení. Pokud nenajdete žádné vhodné rozdělení, můžete tvrdit, že je klasifikace neadekvátní?

Literatura

Levickij, V. V., Kiiko, J. J., Spolnicka, S. V. (1996). Quantitative analysis of verb polysemy in modern German. *Journal of Quantitative Linguistics 3(2)*, 132–135.

2.16 SLOVESA A OSOBY

Problém

Slovesa mohou být rozdělena do skupin různými způsoby. Neexistuje žádné „nejlepší" řešení. Každá klasifikace je podmíněna cílem výzkumu. V tomto problému budeme testovat klasifikaci Scheibmanové (2001).

Hypotéza

„…mohli bychom očekávat větší spoluvýskyt jednotek, jejichž kombinace umožňuje řečníkovi vyjádřit jeho stanovisko, než těch, které to neumožňují (např. podle Benvenistea [1971] by se slovesa označující mentální jevy [slovesa myšlení] měla častěji vyskytovat v 1. os. sg. než v 3. os. sg.)." (Scheibmanová 2001: 65).

Scheibmanová rozděluje slovesa do 10 tříd podle Hallidaye (1994) a uvádí frekvence spojení jednotlivých tříd s gramatickými osobami. Jazykovým materiálem je konverzace. Data jsou uvedena v tabulce 2.16

Levickij a Lučak (2005) vytvořili 20 sémantických podskupin sloves. Viz také Jurčenko (1985), Levin (1998), Sil´nickij (1966).

TABULKA 2.16
Frekvence slovesných tříd v gramatických osobách dle Scheibmanové (2001:65)

Typ slovesa	1.sg.	2.sg.	3.sg.	1.pl.	2.pl.	3.pl.
myšlení	195	110	15	6	0	14
pohybu	24	7	30	1	1	3
existenční	12	6	62	3	0	8
pocitové	19	9	10	2	0	5
materiální	141	90	176	30	2	100
percepční	27	19	6	10	0	2
percepční/rel	0	0	35	0	0	4
posesivní/rel	21	31	29	5	0	16
vztahové	50	41	497	6	2	45
mluvení	128	335	931	66	5	218

Postup

Nejprve proveďte celkový test nezávislosti osoby a typu slovesa. Potom testujte každou buňku zvlášť (udělejte test pro jednotlivé buňky) pro zjištění signifikantních asociací. Ověřte několik dalších hypotéz Scheibmanové na jejích datech. Potom vytvořte výběrový soubor z jiného jazyka a celý postup zopakujte. Ověřte, zda jsou výsledky identické.

Další problémy: v kvantitativní lingvistice je všeobecně známa hypotéza, že pokud jsou některé jednotky „adekvátně“ rozděleny do skupin, pak se ranková frekvenční distribuce těchto jednotek obvykle řídí „správným“

rankovým frekvenčním rozdělením. Testujte, zda data Scheibmanové tuto hypotézu potvrzují.

Je možné uvažovat o asociaci sloves s kategorií osoby za charakteristiku textu podobnou Busemannovu poměru sloves a adjektiv? Můžeme třídit slovesné skupiny podle konceptu „aktivity" nebo podle biologického vývoje života? (Počínaje slovesy bytí až po slovesa duševních stavů, mentálních procesů…) Nebo můžeme provést aposteriorní klasifikaci pro tento problém? Hypotézu by mohlo ujasnit škálování.

Literatura

Benveniste, E. (1971). *Problems in general linguistics.* Coral Gables, Florida: University of Miami Press.

Halliday, M. A. K. (1994). *An introduction to functional grammar.* London: Arnold.

Jurčenko, G. E. (1985). K voprosu o semantičeskoj klassifikacii glagolov anglijskogo jazyka. In: *Grammatičeskaja semantika*. Gorkij: Gorkij University Press, 45–50.

Levickij, V., Lučak, M. (2005). Category of tense and verb semantics in the English language. *Journal of Quantitative Linguistics 12(2–3)*, 212–238.

Levin, B. (1998). *English verb classes and alternations*. Chicago: University of Chicago Press.

Scheibman, J. (2001). Local patterns of subjectivity in person and verb type in American English conversation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 61–89.

Silnickij, G. G. (1966). Semantičeskie klassy glagolov i ich rol´ v tipologičeskoj semasiologii. *Strukturno-tipologičeskie opisanie sovremennych germánských jazykov*, 244–259.

2.17 DISTRIBUCE SLOVNÍCH TŘÍD

Hypotéza

Rankové frekvenční distribuce různých slovních tříd mají stejné pravděpodobnostní rozdělení.

Postup

Spočítejte v textu odděleně různé slovní třídy (substantiva, slovesa, adjektiva, adverbia,…). Pokud se objeví nejednoznačné případy, rozhodněte ad hoc, ke které slovní třídě náleží. Potom aplikujte stejné pravděpodobnostní rozdělení na všechny empirické distribuce, např. Zipfovo useknuté rozdělení zeta $P_x = C/x^a$ $(x = 1,2,3,\ldots,n)$, kde C je normalizační konstanta a $n = x_{max}$. Zkoumejte chování parametru a. Je stejný ve všech případech, nebo se vykytují rozdíly? Porovnejte výsledky s podobnými analýzami textu v jiných jazycích, a to i v případě, že použité slovní třídy nejsou totožné. Pokud je to možné, uspořádejte slovní třídy podle parametru a.

Analyzujte několik jazyků se stejnými slovními třídami, přiřaďte každé slovní třídě pořadí podle parametru a a proveďte srovnávací test rovnoměrnosti uspořádání (viz např. Gibbons 1971). Pokuste se z vašich výsledků vyvodit závěry. Zopakujte výpočty s jinými rozděleními s jedním parametrem, vyvoďte závěry.

Literatura

Gibbons, J. D. (1971). *Nonparametric statistical inference*. New York: McGraw-Hill.

Popescu, I.-I., Vidya, M. N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008). *Word frequency studies*. Berlin, New York: de Gruyter.

3 Kompozita a lexikologie

3.1 STÁŘÍ SLOVA A TENDENCE K TVOŘENÍ KOMPOZIT

Hypotéza

„Čím je slovo starší, tím více kompozit vytváří." (Altmann 1989).

Postup

V tomto případě je nutné pracovat s jednotlivými slovními třídami (konkrétně slovními druhy) odděleně, protože tendence k tvoření složenin je v jednotlivých slovních třídách odlišná. Aby mohla být vytvořena hypotéza, je nutné pracovat s historickým slovníkem. V něm by měl být uveden rok nebo alespoň století prvního výskytu slova v psaných dokumentech. Vytvořte výběrový soubor několika slov stejné slovní třídy, poznamenejte si jejich první výskyt a určete počet jejich kompozit v současném jazyce. Vytvořte graf zaznamenávající empirickou křivku a pokuste se odvodit teoretickou funkci podle nějakých lingvistických předpokladů. Potom pracujte s dalšími slovními třídami. Stejný postup aplikujte na různé jazyky.

Literatura

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10*, 100–107.

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie*. Göttingen: Peust & Gutschmidt.

Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology*. Černivci: RUTA, 177–189.

3.2 KOLOKACE

Postup

„…čím častěji se dva prvky vyskytují v řadě za sebou, tím pevnější bude jejich konstituentní struktura." (Bybee, Hopper 2006: 14). Kolokace mohou být nalezeny testováním koheze dvou slov.

Postup

Vyberte jakékoliv slovo z korpusu a najděte všechna různá slova, která se objeví bezprostředně za ním v dané klauzi. Potom vypočítejte významnost kolokace za použití hypergeometrického rozdělení a Poissonova rozdělení. Vypočítejte podmíněnou pravděpodobnost následujícího slova. Vyhodnoťte kolokaci určením pravěpodobnostní hranice signifikantnosti. (viz část 4.1, „Asociační graf textu").

Literatura

Bisht, R. K., Dhami, H. S., Tiwari, N. (2006). An evaluation of different statistical techniques of collocation extraction using a probability measure to word combinations. *Journal of Quantitative Linguistics 13(2–3)*, 161–175.

Bybee, J., Hopper, P. (eds.) (2001). *Frequency and the emergence of linguistic Structure*. Amsterdam: J. Benjamins.

Levickij, V. (2005). Lexikalische Kombinierbarkeit. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international Handbook*. Berlin, New York: de Gruyter, 464–470.

Levickij, V. V., Zadorožna, I. (2007). Die Stärkemessung des Zusammenhangs zwischen den Komponenten der Phraseologismen. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text.* Berlin, New York: de Gruyter, 399–406.

Lin, D. (1998). Extracting collocations from text corpora. *First Workshop on Computational Terminology.* Montreal.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics 19(1)*, 143–177.

3.3 DÉLKA KOMPOZITA A DÉLKA KOMPONENTU

Hypotéza

„Čím delší je kompozitum, tím kratší jsou jeho komponenty." (Altmann 1989).

Postup

Tato hypotéza vyplývá z Menzerathova zákona: čím delší je konstrukt, tím kratší jsou jeho konstituenty. Testování tohoto zákona je velice jednoduché: vytvořte (náhodně) seznam kompozit ze slovníku nebo korpusu. Použijte dva druhy měření délky komponentů: (a) v počtu fonémů, (b) v počtu slabik. Délka kompozita je měřena počtem jeho komponentů. Spočítejte délku každého kompozita a průměrnou délku jeho komponentů. Pokud hypotéza platí, výsledkem budou dvě monotónní funkce <*délka kompozita, průměrná délka komponentu*>. Bohužel kromě maďarštiny nebo němčiny jsou kompozita s více komponenty v jazycích spíše výjimkou. V případě potřeby tedy použijte speciální slovník, který obsahuje dlouhé složeniny.

Literatura

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10*, 100–107.

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie*. Göttingen: Peust & Gutschmidt.

Cramer, I. (2005). Das Menzerathsche Gesetz. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 659–688.

Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology*. Černivci: RUTA, 177–189.

3.4 DÉLKA KOMPOZIT A JEJICH KOTEXTUALITA

Hypotéza

„Čím delší je kompozitum, tím menší je jeho kotextualita." (Altmann 1989).

Postup

Vytvořte náhodný výběrový soubor kompozit z korpusu. Zjistěte délku jednotlivých kompozit (v počtu komponentů). Potom vypočtěte jejich kotextualitu jedním ze způsobů uvedených v předchozí hypotéze. Pokuste se vytvořit relaci <*délka kompozita, rozsah kotextuality*> pomocí empirické funkce i teoretického zdůvodnění.

Literatura

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10*, 100–107.

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie*. Göttingen: Peust & Gutschmidt.

Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology*. Černivci: RUTA, 177–189.

3.5 DÉLKA KOMPOZIT A POLYSÉMIE

Hypotéza

„Čím delší kompozitum je, tím má méně významů (v průměru)." (Altmann 1989).

Postup

Tento problém je formálně stejný jako problém vztahu mezi délkou slova a polysémií. Rozdíl je v tom, že délka kompozita je měřena počtem jeho komponentů. Vzhledem k tomu, že skládání slov naplňuje požadavek[1] specifikace, hypotéza musí platit. Pokuste se nalézt relaci <*počet komponentů, polysémie kompozit*>. Začněte od proporcí a pracujte s průměry, jinak může být výsledkem nesmírný rozptyl. Jakmile bude stanoveno několik hypotéz o kompozitech, pokuste se vytvořit kontrolní cyklus obsahující vztah *polysémie kompozit = f (počet komponentů, jiná proměnná).*

Literatura

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10*, 100–107.

Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology*. Černivci: RUTA, 177–189.

3.6 DÉLKA KOMPOZIT A SÉMANTICKÁ SHODA

Hypotéza

Čím delší je kompozitum, tím větší je sémantická shoda s jeho komponenty.

Postup

Délka kompozita je vyjádřena počtem jeho komponentů (nikoliv slabik). Navažte na předchozí problém, ale tentokrát vytvořte náhodný výběrový

1 Pozn. překladatele: je míněn „požadavek" (requirement), tak jak se chápe v synergetickém modelu jazyka.

soubor kompozit o různých délkách. Potom stejně jako u předchozího problému proveďte dva způsoby výpočtu.

(1) Vypočítejte průměrnou shodu kompozit různých délek a vytvořte relace <*délka, shoda*>.

(2) Vezměte minimální shody u každé třídy délek a vytvořte stejný vztah. Pokud je to možné, zkoumejte jazyky s delšími složeninami a případně použijte odborné slovníky.

Literatura

Laws in Quantitative Linguistics. [online]. Dostupné z: http://lql.uni-trier.de/index.php/Main_Page

3.7 KOMPOZITA A SÉMANTICKÁ SHODA

Hypotéza

„Počet kompozit v jazyce klesá úměrně s mírou sémantické shody komponentů s daným kompozitem." (Altmann 1989).

Postup

Nejprve vytvořte metodu pro měření sémantické shody mezi kompozity a jejich komponenty. Například v anglickém slově *hangover* (česky kocovina) není žádná významová shoda mezi jednolivými částmi (*hang, over*) a kompozitem *hangover*. Významová shoda v německých slovech *Kindergarten* nebo *Baumschule* je naopak zřejmá: alespoň jeden komponent vyjadřuje samostatně část významu kompozita. Stejná situace je patrná u anglického slova *bookseller*, u kterého se projevuje vysoká sémantická shoda jednotlivých částí a kompozita. Rozlišujte kompozita také podle stupně koheze (nebo typu). Opatřete si výběrový soubor kompozit a vytvořte distribuci

jejich sémantické shody nebo zkuste vytvořit relaci <*sémantická shoda, počet složenin*> či naopak.

Literatura

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10*, 100–107.

Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology*. Černivci: RUTA, 177–189.

Laws in Quantitative Linguistics. [online]. Dostupné z: http://lql.uni-trier.de/index.php/Main_Page

3.8 SKLÁDÁNÍ SLOV A ASOCIACE

Hypotéza

S růstem počtu asociací daného slova roste počet kompozit, které dané slovo vytváří.

Postup

Použijte slovník asociací, který je dostupný v mnoha jazycích. Opatřete si náhodný výběrový soubor slov a uveďte počet asociací (typů). Ignorujte četnost jednotlivých asociací. Potom si vezměte slovník kompozit a pro každé slovo v souboru spočítejte počet kompozit, které dané slovo vytváří.

Pokud hypotéza platí, relace <*počet významů, počet složených slov*> bude mít formu monotónní rostoucí funkce. Vytvořte graf, aplikujte vhodnou teoretickou funkci na empirickou distribuci a zdůvodněte ji na základě argumentů proporcionality.

Literatura

žádná

3.9 SKLÁDÁNÍ SLOV A EMOCIONALITA

Hypotéza

S rostoucím emocionálním významem slova roste počet kompozit, které dané slovo vytváří.

Postup

Slovní emocionalita může být měřena Osgoodovou metodou, dotazníkovou formou nebo na základě psycholingvistické literatury. Vezměte 100 slov a změřte jejich emocionalitu (jakéhokoliv druhu). Následně za pomocí slovníku spočítejte počet kompozit, které dané slovo vytváří. Vyneste křivku <*emocionalita, počet kompozit*> vyjadřující tuto závislost. Potom odvoďte adekvátní funkci na základě argumentů proporcionality.

Literatura

žádná

3.10 KOTEXTUALITA A TENDENCE K TVOŘENÍ KOMPOZIT

Hypotéza

„Čím větší je kotextualita slova, tím více kompozit dané slovo vytváří." (Altmann 1989).

Postup

Kotextualita může být měřena dvěma různými způsoby: (a) na základě počtu textů (v korpusu), v nichž se dané slovo objeví, nebo (b) na základě počtu různých slov vyskytujících se v sousedství daného slova, tj. na základě počtu kontextů. S rostoucí frekvencí daného slova roste potřeba jej

specifikovat. Skládání slov je jedna z možností specifikace slova, proto očekáváme, že frekventovaná slova budou mít tendenci častěji tvořit kompozita. Při testování hypotézy rozlišujte jednotlivé slovní třídy, vezměte například jen substantiva nebo slovesa. Jakmile vypočítáte kotextualitu slov, zjistěte, v kolika kompozitech se každé slovo vyskytuje. Potom se pokuste vysledovat tendenci <*kotextualita, počet kompozit*>. Vytvořte graf, jež bude zobrazovat monotónní rostoucí tendenci. Vyneste empirickou křivku a pokuste se ji zdůvodnit za pomoci nějakých teoretických předpokladů.

Literatura

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10*, 100–107.

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie*. Göttingen: Peust & Gutschmidt.

Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology*. Černivci: RUTA, 177–189.

3.11 DISORTATIVITA SKLÁDÁNÍ SLOV

Hypotéza

Technika skládání slov je disortativní.

Postup

Nechť je počet různých slov, se kterými určité slovní tvary tvoří kompozita, jejich stupněm (tento temín je převzat z teorie grafů, srov. anglický termín „degree“). Pokud slova s vysokým stupněm inklinují k tvoření kompozit se slovy s vysokým stupněm, skládání slov je asortativní. Pokud slova s vysokým stupněm inklinují k tvoření kompozit se slovy s nízkým stupněm, skládání slov je disortativní. V ostatních případech je neutrální.

Vytvořte velký výběrový soubor kompozit (pokud možno všechna kompozita ze slovníku) a vypočítejte stupeň každého komponentu. Potom pro všechny komponenty, které mají stejný stupeň, vypočítejte průměrný stupeň komponentů tvořících s nimi kompozita. Vytvořte relaci <*stupeň komponentu, průměrný stupeň sousedních komponentů*>. Pokud je skládání slov disortativní, výsledkem bude monotónní klesající funkce. Pokud je skládání slov asortativní, výsledkem bude monotónní rostoucí funkce. V případě neutrality skládání slov bude výsledkem rovná přímka.

Tento problém může být řešen také s kombinacemi fonémů (bigramy).

Literatura

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters 89(20)*, article: 208701.

Tamaoka, K., Meyer, P., Makioka, S., Altmann, G. (2008). On the dynamics of compounding of Japanese kanji with common and proper nouns. *Journal of Quantitative Linguistics 15(2),* 165–153.

3.12 DISTRIBUCE DÉLKY KOMPOZIT

Hypotéza

„Počet kompozit klesá s jejich rostoucí délkou." (Altmann 1989).

Postup

Hypotéza říká, že distribuce kompozit je jednoduchá monotónní klesající funkce. Vytvořte náhodný výběrový soubor kompozit. Délka kompozita je měřena počtem jeho komponentů. Zkonstruujte empirickou distribuci délek kompozit a najděte buď odpovídající teoretické rozdělení, nebo funkci vyjadřující vzah <*délka kompozita, počet kompozit dané délky*>. Pokud platí Mezerathův zákon, vhodným modelem by měla být zeta distribuce

nebo funkce zeta. Pokuste se testovat hypotézu v různých jazycích a problém zobecnit. Vezměte v potaz, že skládání slov je výrazem požadavku specifikace.

Literatura

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10*, 100–107.

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie*. Göttingen: Peust & Gutschmidt.

Fan, F., Altmann, G. (2007). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds), *Problems of typological and quantitative lexicology*. Černivci: RUTA, 177–189.

3.13 DISTRIBUCE SYNONYM

Hypotéza

Distribuce synonym ve slovníku je pravidelná.

Postup

Použijte slovník synonym. Vytvořte systematicky výběrový soubor slov, např. vezměte každé poslední slovo na stránce. Spočítejte, kolik slov má přesně $x = 1,2,3,\ldots$ synonym. Na získaná data aplikujte známé modely.

Literatura

Uhlířová, L. (2001). Kolik je v češtině synonym? (K dynamické stabilitě v systému lexikálních synonym). In: Ondrejovič, S., Považaj, M. (eds.), *Lexicographica ´99*. Bratislava: Veda, 237–250.

Wimmer, G., Altmann, G. (2001). Two hypotheses on synonymy. In: Ondrejovič, S., Považaj, M. (eds.), *Lexicographica ´99*. Bratislava: Veda, 218–225.

3.14 NÁRŮST POČTU PŘEJATÝCH SLOV

Hypotéza

Počet přejatých slov se ve všech jazycích zvyšuje podle Piotrowského zákona.

Postup

Použijte historický slovník daného jazyka a spočítejte počet cizích slov, které se v něm vyskytují. U každého slova si pak poznamenejte (přibližně) rok, kdy do sledovaného jazyka proniklo.

Popřípadě zvolte druhý způsob: v časopisu nebo novinách zjistěte počet nových anglických slov, která byla do daného jazyka přejata po roce 1950. Poznamenejte si pouze první rok, kdy bylo dané slovo přejato.

Testujte hypotézu, podle níž se kumulativní počet cizích slov řídí Piotrowského zákonem

$$y_t = \frac{c}{1 + ae^{-bt}} \quad ,$$

kde y_t je počet cizích slov v čase t, C je asymptota a a, b jsou parametry. Pro analýzu můžete také použít speciální slovníky nebo katalogy obchodních domů.

Literatura

Altmann, G. (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, K.-H., Kohlhase, J. (eds.), *Exakte Sprachwandelforschung*. Göttingen: edition herodot, 54–90.

Best, K.-H. (2003). Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur reichweite des Piotrowski-Gesetzes. *Glottometrics* 6, 9–34.

Best, K.-H. (2006). Deutsche Entlehnungen im Englischen. *Glottometrics 13*, 66–72.

Best, K.-H. (2004). Zur Ausbreitung von Wörtern arabischer Herkunft im Deutschen. *Glottometrics 8*, 75–78.

Best, K.-H. (2005). Turzismen im Deutschen. *Glottometrics 11*, 56–63.

Best, K.-H., Altmann, G. (1986). Untersuchungen zur Gesetzmäßigkeit von Entlehnungsprozessen im Deutschen. *Folia Linguistica Historica 7*, 31–41.

Körner, H. (2004). Zur Entwicklung des deutschen (Lehn-) Wortschatzes. *Glottometrics 7*, 25–49.

3.15 LEXIKÁLNÍ ŘETĚZCE

Problém

Popište aspekty hyperonymické struktury lexika angličtiny (nebo jiného jazyka).

Postup

Hyperonymum lexému A je jiný lexém určující třídu, do které A náleží. Například *nábytek* je hyperonymem *křesla*, *budova* je hyperonymem *mrakodrapu*. Hyperonymum je obvykle uvedeno v definici významu ve výkladovém slovníku. Vezměte do úvahy substantiva, která tvoří hyperonymické řetězce. Wordnet poskytuje hyperonymické řetězce pro angličtinu, pro ostatní jazyky musí být nově vytvořeny[2]. Při vytváření hyperonymických řetězců věnujte pozornost následujícím bodům:

(1) Neuvažujte žádné jiné vztahy než příslušnost k téže třídě, tzn. nezapočítavejte vztahy ve smyslu „část něčeho", jako hlava = část těla, motor = část auta (tělo není hyperonymem hlavy a auto není hyperonymem motoru).

(2) Uvažujte pouze první, hlavní význam substantiva. Pokud má více významů, utvořte řetězec ke každému zvlášť.

2 Pozn. překladatele: v dnešní době jsou dostupné Wordnety pro celou řadu jazyků, včetně češtiny.

(3) Vyhněte se cirkularitě (která se objevuje i ve Wordnetu).

(4) Vybírejte hyperonyma značně vysoké obecnosti či abstraktnosti, např. *entita*, *systém*, *bytí*, *věc* apod., ale vyřazujte definice typu *něco*, *co*.

(5) Nevyřazujte abstraktní substantiva.

(6) Pokud se nějaké substantivum vyskytne v jakémkoli řetězci jako hyperonymum, nezařazujte jej již do množiny základních lexémů. Jakmile budete mít data připravena, proveďte následující kroky.

   (a) Pokuste se najít distribuci délky lexikálních řetězců jak empiricky, tak teoreticky.

   (b) Seřaďte řetězce tak, aby byl základní lexém na první úrovni. Spočítejte průměrnou délku lexémů na první, druhé, třetí,… atd. úrovni. Pokuste se najít nějaký trend.

   (c) Sledujte počet (podíl) jednomorfémových slov na první, druhé, třetí,… atd. úrovni. Pokuste se najít nějaký trend.

   (d) Spočítejte, kolik různých slov (typů) je na první, druhé, třetí,… atd. úrovni a sledujte, zda výskyt typů pravidelné klesá.

   (e) Seřaďte řetězce tak, aby nejvyšší hyperonyma (konce řetězce) byla na první úrovni. Následně proveďte úkoly (b), (c) a (d).

Literatura

Hammerl, R. (1987). Untersuchungen zur mathematischen Beschreibung des Martingesetzes der Abstraktionsebenen. *Glottometrika 8*, 113–129.

Hammerl, R. (1989). Neue Perspektiven der sprachlichen Synergetik: Begriffsstrukturen – kognitive Netze. *Glottometrika 10*, 129–140.

Hammerl, R. (1989). Untersuchung struktureller Eigenschaften von Begriffsnetzen. *Glottometrika 10*, 141–154.

Sambor, J. (2005). Lexical networks. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 447–458.

Sambor, J., Hammerl, R. (eds.) (1991). *Definitionsfolgen und Lexemnetze*. Band I. Lüdenscheid: RAM.

Schierholz, S. (1989). Kritische Aspekte zum Martinschen Gesetz. *Glottometrika 10*, 108–128.

3.16 LEXIKÁLNÍ SÍTĚ

Problém

Vytvořte definiční řetězce obsahující hyperonyma pro všechny významy daného lexému. Následně vytvořte z těchto řetězců lexikální síť. Zkoumejte vlastnosti této sítě.

Postup

Vyberte z výkladového slovníku náhodně 100 substantiv. Slova s jedním významem budou mít jednoduchý řetězec, spojující je s lexémem s nejobecnějším významem. V případě polysémních slov však budou různé cesty k nejobecnějšímu lexému nebo lexémům. Z těchto řetězců může být vytvořen orientovaný graf, který má několik vlastností. Vyhodnoťte alespoň následující vlastnosti:

(1) počet výrazů v grafu (= počet vrcholů) a jejich distribuci v jazyce (nejméně 100 substantiv),

(2) šířka grafů definovaná Hammerlem (1989),

(3) počet větví a jejich distribuce,

(4) počet koncových lexémů a jejich distribuce,

(5) průměrná délka jednotlivých větví (cest),

(6) síla sémantických vztahů mezi lexémy v síti,

(7) produktivita lexémů atd.

Použijte všechny prostředky z teorie grafů k charakterizování takto konstruovaných lexikálních sítí v jazyce.

Navrhněte metodu pro podobnou analýzu sloves a adjektiv. Porovnejte více jazyků.

Tato oblast výzkumu je v současnosti nedostatečně prozkoumána, proto je nezbytné provést další šetření.

Literatura

Hammerl, R. (1989). Untersuchung struktureller Eigenschaften von Begriffsnetzen. *Glottometrika 10*, 141–154.

Kisro-Völker, S. (1984). On the measurement of abstractness in lexicon. *Glottometrika* 6, 139–151.

Sambor, J. (2005). Lexical networks. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 447–458.

Sambor, J., Hammerl, R. (eds.) (1991). *Definitionsfolgen und Lexemnetze*. Band 1. Lüdenscheid: RAM.

Skorochod´ko, E. F. (1981). *Semantische Relationen in der Lexik und in Texten*. Bochum: Brockmeyer.

3.17 DÉLKA KMENU A TENDENCE K TVOŘENÍ KOMPOZIT

Hypotéza

„Čím kratší je nějaké slovo, tím častěji se vyskytuje v kompozitech." (Altmann 1989).

Postup

Vytvořte náhodný výběrový soubor slovních kmenů, nejlépe v rámci jednoho slovního druhu. Potom vyhledejte všechna kompozita, která tyto kmeny obsahují. Poznamenejte si pozici kmenu uvnitř kompozita. Vytvořte graf, abyste mohli sledovat tendenci: čím delší je kmen, tím méně kompozit vytváří. Následně se pokuste prostřednictvím argumetu proporcionality odvodit relaci <*délka kmenu, počet kompozit*>. Mohou být použity i již publikované zdroje.

Literatura

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10*, 100–107.

Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology*. Černivci: RUTA, 177–189.

3.18 DÉLKA SLOV A SYNONYMIE

Hypotéza

Čím delší je slovo, tím méně má synonym.

Postup

Použijte slovník synonym a vytvořte rozsáhlý náhodný výběrový soubor slov. Následně zjistěte počet jejich synonym. Pokuste se najít druh závislosti počtu synonym na délce slova. Pro každou délku slova je potřeba počítat průměrný počet synonym.

Literatura

Wimmer, G., Altmann, G. (2001). Two hypotheses on synonymy. In: Ondrejovič, S., Považaj, M. (eds.), *Lexicographica´99*. Bratislava: Veda, 218–225.

4 Textologie

4.1 ASOCIAČNÍ GRAF TEXTU

Problém

Slova mohou být v textu asociována skrytě (nikoliv vytvářením kolokací). Vytvořte graf asociací a vyhodnoťte vlastnosti tohoto grafu.

Postup

Spočítejte absolutní frekvence substantiv, sloves a adjektiv v jednom celém textu. Následně spočítejte počet vět N v textu. Vezměte první dvě slova z vašeho seznamu substantiv, sloves a adjektiv. Frekvence prvních slov označte M a frekvence druhých slov n. Potom určete počet vět x, kde se obě slova objevila zároveň. Aby bylo možné testovat asociační sílu, proveďte následující výpočet: pokud $x > Mn/N$, pak

$$P(X \geq x) = \sum_{j=x}^{min[n,M]} \frac{\binom{M}{j}\binom{N-M}{n-j}}{\binom{N}{n}} \ ,$$

Zvolte hladinu významnosti $\alpha = 0{,}05$. Pokud je P menší než α, můžeme tvrdit, že mezi dvěma slovy existuje asociace.

Proveďte test pro všechny dvojice slov z vašeho seznamu. Vytvořte graf slovních asociací a prozkoumejte jeho vlastnosti. Aplikujte tento postup na texty různých žánrů a jazyků. Stanovte hypotézu o asociační struktuře textu.

Literatura

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Popescu, I.-I., Vidya, M. N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008). *Word frequency studies*. Berlin, New York: de Gruyter.

4.2 SHLUKOVÁNÍ AUTOSÉMANTIK V DANÝCH INTERVALECH

Problém

Rozdělením rankové frekvenční distribuce na intervaly o velikosti h (h označuje h-bod) získáme exponenciálně vzrůstající počet autosémantik v po sobě jdoucích intervalech. Testujte toto tvrzení na datech z různých textů.

Postup

Vytvořte rankovou frekvenční distribuci slov u zvoleného textu a vypočítejte h-bod. Rozdělte tuto distribuci v krocích o velikosti h od prvního ranku až k nejvyššímu. Spočítejte autosémantika v každém intervalu z tabulky, kde budou intervaly označeny jako 1., 2.,…, tím změníte měřítko distribuce. Získate tak rostoucí funkci. Pokuste se na získaná data aplikovat funkci

$$y = a[1 - \exp(-kx)] .$$

Popescu et al. (2008) definovali dvě textové charakteristiky: *autosémantickou kompaktnost* $AC = ak$, kde a a k jsou parametry výše uvedené funkce, a shlukování autosémantik v daných intervalech (*autosemantic pace filling*) $APF = a/h$. Vypočítejte tyto dva indikátory pro mnoho textů. Abyste mohli testovat rozdíly APF a AC mezi dvěma texty, použijte testy uvedené v literatuře (Popescu et al. 2008).

Ke klasifikaci textu použijte souřadnice <*1/k, a*> získané z výše uvedené funkce. Následně proveďte lineární diskriminační analýzu nebo použijte nějakou moderní taxonomickou metodu. Získané výsledky interpretujte.

Literatura

Popescu, I.-I., Vidya, M. N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008). *Word frequency studies*. Berlin, New York: de Gruyter

4.3 CARROLLŮV VEKTOR

Problém

V téměř zapomenutém článku navrhl Carroll (1960) několik potenciálních charakteristik textu. Vytvořte dvojice těchto charakteristik a testujte jejich vzájemnou nezávislost.

Postup

Změřte objektivní a subjektivní charakteristiky několika textů. Pokuste se najít argumenty, pomocí nichž by se dala zdůvodnit vzájemná závislost některých párů těchto charakteristik. Pokud je to možné, odvoďte závislost pomocí diferenciální rovnice. Spojujte jednotlivé charakteristiky krok za krokem k rostoucím korelačním sadám, abyste získali kontrolní okruh podobný Köhlerovu (1986).

Použijte nezávislé vlastnosti k charakteristice textů a kontrolní okruh/okruhy k vytvoření základní teorie. Neomezujte svůj výzkum pouze na prózu. Použijte také charakteristiky prezentované v Tuldava (1995: 93–108). Vytvořte vlastní škálu pro vyhodnocování subjektivních charakteristik.

Literatura

Carroll, J. B. (1960). Vectors of prose style. In: Sebeok, T. A. (ed.), *Style in language*. Cambridge: The MIT Press, 283–292.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Tuldava, J. (1995). *Methods in quantitative linguistics*. Trier: Wissenschaftlicher Verlag.

Tuldava, J. (2005). Stylistics, author identification. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 368–387.

4.4 ALTERNATIVNÍ TYPE-TOKEN INDEX

Problém

Ejiri a Smith (1993) navrhli alternativní type-token míru textu ve formě $G = \log(N/L)/\{\log(N)\text{-}1\}$, kde N = délka textu, L = slovník.

Postup

Popište nebo jen okomentujte metodologickou a statistickou povahu tohoto indexu. Co tento index říká? V jakém intervalu se pohybují výsledky? Jaké jsou jeho statistické odhady a variance? Jak je možné porovnávat dva texty prostřednictvím tohoto indexu? Zamyslete se nad tvořením indexu obecně. Jaké vlastnosti musí index mít?

Literatura

Ejiri, K., Smith, A. E. (1993). Proposal for a new ‚Constraint measure' for text. In: Köhler, R., Rieger, B. B. (eds.), *Contributions to quantitative linguistics*. Dordrecht: Kluwer, 195–211.

Galtung, J. (1967). *Theory and methods of social research*. Oslo: Universitetsforlaget.

4.5 KOTEXTUALITA A FREKVENCE

Hypotéza

Čím větší je kotextualita nějaké jednotky, tím vyšší je její frekvence.

Postup

Zvažte různé způsoby pojetí kotextuality. Vzhledem k fonému může být kotextualita určena počtem ostatních fonémů, které se mohou objevit bezprostředně vedle daného fonému, počtem různých typů slabik nebo počtem různých slov, ve kterých se daný foném vyskytne. Vzhledem ke slabice je kotextualita určena počtem různých slovních tvarů, ve kterých se daná slabika nachází. Vzhledem ke slovu je určena počtem různých textů, ve kterých se dané slovo vyskytuje. Hypotéza předpokládá, že silná kotextualita má za následek vysokou frekvenci jednotek, ačkoliv to však ve skutečnosti nemusí vždy platit. Tato hypotéza je součástí Köhlerova samoregulačního cyklu. Testujte hypotézu jakýmkoliv způsobem. Zvažte také možnost změny směru závislosti.

Určete kotextualitu a frekvence slov z dlouhého textu nebo korpusu. Pokud hypotéza platí, bude mít podobu mocninné funkce. Pokud hypotéza neplatí, zkuste ji modifikovat.

Literatura

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Tamaoka, K., Makioka, S. (2004). Frequency of occurrence for units of phonemes, morae, and syllables appearing in a lexical corpus of a Japanese newspaper. *Behavior Research Methods, Instruments, & Computers 36(3)*, 531–547.

4.6 VZDÁLENOSTI MEZI STEJNĚ DLOUHÝMI VĚTAMI

Hypotéza

Vzdálenosti mezi stejně dlouhými větami v textu se řídí Zipf-Aleksejevovým rozdělením (Hřebíček 2000: 36nn).

Postup

Definujte různými způsoby vzdálenost mezi stejně dlouhými větami v dlouhém textu. Nejjednodušší způsob je spočítat počet odlišně dlouhých vět mezi dvěma stejně dlouhými větami. Vytvořte distribuci vzdáleností a testujte, zda Zipf-Aleksejevovo rozdělení odpovídá výsledkům.

Vezměte v úvahu ostatní textové jednotky a zkoumejte více jazyků, abyste mohli hypotézu relevantněji potvrdit, nebo vyvrátit.

Pokud Zipf-Aleksejevovo rozdělení není vhodným modelem, pokuste se najít jiné řešení, které by potvrzovalo hypotézu.

Literatura

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

4.7 VZDÁLENOSTI MEZI LEXÉMY

Hypotéza

Vzdálenosti mezi výskyty stejného lexému se řídí mocninným zákonem (Hřebíček 2000:32nn).

Postup

Definujte vzdálenost mezi identickými lexémy v dlouhém lemmatizovaném textu počtem lemmat nebo vět ležících mezi nimi. Výzkum můžete omezit i na jediný lexém. Zjistěte vzdálenosti mezi identickými lexémy a vytvořte jejich distribuci. Pokud se některé vzdálenosti nebudou vyskytovat v dostatečném množství, bude nutné je sloučit s jinými do jedné třídy vzdálenosti. Transformujte takto vytvořené třídy na $x = 1, 2, 3,\ldots$ (jednoduchým přejmenováním). Potom testujte, zda se relace <*x, počet vzdáleností velikosti x*> řídí mocninným zákonem.

Stejný postup může být aplikován na slovní tvary nebo jiné vhodné jednotky (slabiky, morfémy). Pokuste se najít souvislost se Skinnerovou hypotézou (srov. část 4.21, „Fonetická agregace“).

Literatura

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

4.8 EUFONIE

Problém

Eufonie může být v textu dosaženo buď na základě mimořádné frekvence jednotlivých fonémů, prostřednictvím jejich spojování, nebo jejich umístěním v textu (např. rým). Pokuste se vytvořit způsob měření eufonie, popř. alespoň vytvořte pracovní definici.

Postup

Zjistěte relativní četnosti fonémů v jazyce z vybraného souboru nepoetických textů (použijte např. korpus, prozaické texty), p_i bude vyjadřovat relativní četnost fonému *i* a ξ bude náhodná proměnná reprezentující počet výskytů fonému *i*. Nechť n je počet fonémů ve verši (řádku) a α zvolená

hladina významnosti, např. $\alpha = 0{,}05$. Počítejte první foném na řádku. Pokud bude jeho frekvence f_i vyšší než np_i, potom spočítejte kumulativní pravděpodobnost

$$P(\xi \geq f_i) = \sum_{f_i}^{n} \binom{n}{x} p_i^x q_i^{n-x} \ ,$$

kde $q_i = 1 - p_i$. Eufonická váha fonému i v daném řádku je definována jako

$$E(i) = \begin{cases} 100[a - P(\xi \geq f_i)], \ pokud \ a > P(\xi \geq f_i) \\ 0 \ v \ opačném \ případě \end{cases} .$$

Nechť je E je množina fonémů, jejichž $E(i) > 0$ a $k = |E|$. Následně může být eufonie řádku definována jako

$$E(řádek) = \frac{100}{k} \sum_{i \in E} [\alpha - P(\xi \geq f_i)] \ ,$$

tj. průměrná eufonie všech eufonických fonémů. Nechť je N je počet řádků v básni. Následně může být eufonická hodnota celé básně definována jako

$$E(báseň) = \frac{1}{N} \sum_{j=1}^{N} E(řádek_j) \ .$$

Pokuste se provést následující kroky:

(1) Analyzujte báseň výše popsaným způsobem.

(2) Zjistěte, zda má eufonie specifický průběh od začátku do konce básně.

(3) Analyzujte vývoj eufonie jednoho autora nebo jednoho jazyka.

(4) Navrhněte další způsoby měření eufonie.

(5) Zjistěte, zda existuje vztah mezi eufonií a významem básně či dalšími vlastnostmi textu.

Literatura

Altmann, G. (1966). The measurement of euphony. In: *Teorie verše I*. Brno: Universita J. E. Purkyně, 259–261.

Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S. (2003). *Úvod do analýzy textov.* Bratislava: Veda.

4.9 HIRSCHŮV-POPESCŮV BOD

Problém

Hirschův-Popescův bod je takový bod, pro nějž platí $x = f(x)$ v rankové frekvenční distribuci nebo ve frekvenčním spektru. Nechť F_h je kumulativní frekvence do h-bodu, tj. $F_h\ (X \leq h)$. Zodpovězte následující otázky:

(1) Koreluje h-bod s entropií a indexem opakování?

(2) Které z mnoha indexů slovního bohatství textu korelují s h-bodem nebo F_h?

(3) Je h-bod závislý na délce textu?

(4) Existuje rozdíl v h-bodu mezi žánry?

(5) Charakterizuje h-bod různé autory?

(6) Existují rozdíly mezi h-body jednotlivých jazyků?

Postup

Přečtěte si část 7.6, „Popescův typologický indikátor *a*“, v níž je výpočet h-bodu, a část 8.10, „Index opakování a entropie“. Vyberte jeden text, vypočítejte všechny uvedené indexy a vyřešte dané problémy.

Pokračování (pro matematiky)

Pokuste se odvodit *h*-bod pro některé diskrétní pravděpodobnostní rozdělení používané pro modely rankové frekvenční distribuce. Pokud je výpočet obtížný, přidejte aproximaci pomocí integrálů, rozvoje řad atd.

Literatura

Hirsch, J. E. (2005). *An index to quantify an individual´s scientific research output.* Dostupné z: *http://arxiv.org/PS_cache/physics/pdf/0508/0508025.pdf*

Mačutek, J., Popescu, I.-I., Altmann, G. (2007). Confidence intervals and tests fort he h-point and related text characteristics. *Glottometrics 15*, 42–52.

Popescu, I.-I. (2007). The ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text*. Berlin, New York: de Gruyter, 553–562.

Popescu, I.-I., Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics 13*, 23–46.

Popescu, I.-I., Altmann, G. (2007). Writer´s view of text generation. *Glottometrics 13*, 71–81.

4.10 HREBY

Problém

Jazyková jednotka *hreb* byla pojmenována po Luďku Hřebíčkovi, který ji definoval (on ji nazývá *agregát*). *Hreb* je množina morfémů, slov, frází, klauzí nebo vět, které sdílejí stejný významový prvek.

Hreb je dobře definovatelná jednotka, která se řídí všemi zákony a tendencemi výstavby textu. Testujte některé níže uvedené hypotézy.

Postup

(1) Zaměřte se na text jako na sekvenci morfémů. Označte každý jednotlivý význam morfému jiným číslem. Nahraďte jednotlivé morfémy těmito čísly. Takto získáte sekvence čísel. Zkoumejte tato čísla jako jednotky a popište chování morfémů-*hrebů*.

(2) Zaměřte se na text jako na sekvenci slov. Proveďte se slovy-*hreby* to samé jako v případě morfémů-*hrebů*. Je vhodné vynechat některé třídy slov, např. spojky, předložky, některé číslovky a členy.

(3) Zaměřte se na text jako na sekvenci frází a proveďte výše uvedený postup. Pro získání spolehlivých výsledků by měl být analyzován dostatečně dlouhý text.

(4) Přiřaďte všechny věty obsahující stejný referent ke stejnému *hrebu*. V tomto případě je *hrebem* množina vět, které obsahují *něco společného* (společnou jednotku nebo referenci). Každá věta v textu může patřit do několika různých *hrebů*. Takto mohou být vytvořeny dvě různé množiny (jedna tvořená větami a jedna tvořená *hreby*).

Proveďte následující úkoly. (a) Vytvořte graf obsahující dvě různé skupiny prvků – jednu skupinu bude tvořit množina vět, druhou množina *hrebů* – a spočítejte jeho charakteristiky. (b) Potom vytvořte graf, v němž budou hranou spojeny věty obsahující společný referent. (c) Testujte hypotézu „čím více je vět v jednom *hrebu*, tím jsou tyto věty kratší", což je ve shodě s Menzerathovým zákonem. (d) Se všemi druhy *hrebů* proveďte běžnou denotativní analýzu: (i) vytvořte distribuci velikosti *hrebů*, (ii) vypočítejte difúznost *hrebů* v textu, (iii) vypočítejte kompaktnost textu, (iv) vytvořte graf poziční koincidence *hrebů,* (v) vypočítejte koncentraci textu, (vi) vypočítejte spojitost textu, (vii) spočítejte vzdálenosti mezi *hreby*, (viii) určete skupiny *hrebů* se stejnými vlastnostmi atd.

Zkuste definovat jiné typy *hrebů*. Sledujte více vlastností příslušných grafů a interpretujte je lingvisticky.

Literatura

Hřebíček, L. (1993). Text as a construct of aggregations. In: Köhler, R., Rieger, B. B. (eds.), *Contributions to quantitative linguistics*. Dordrecht: Kluwer, 33–39.

Hřebíček, L. (1995). *Text levels. Language constructs, constituents and the Menzerath-Altmann law*. Trier: Wissenschaftlicher Verlag.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute.

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

Köhler, R., Naumann, S. (2007). Quantitative analysis of co-reference structure in text. In: Grzybek, P., Köhler, R. (eds), *Exact method in the study of language and text*. Berlin, New York: de Gruyter, 317–329.

West, D. B. (2001). *Introduction to graph theory*. Second edition. Upper Saddle River, NJ: Prentice Hall.

Ziegler, A. (2005). Denotative Textanalyse. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, 423–447.

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse. Ein textlinguistisches Arbeitsbuch*. Wien: Edition Prasens.

4.11 HURSTŮV EXPONENT

Problém

Prostřednictvím Hurstova exponentu charakterizujte chování délkových sekvencí.

Postup

Přepište text jako sekvenci délek. Použijte různé jednotky jako délku morfémů, slov, vět atd. (všechny měřeny různými způsoby). Stanovte následující hodnoty:

i = pozice v sekvenci,

x_i = délka jednotky v pozici i,

$R_i = max\, x_i - min\, x_i$,

R_i je rozsah, rozdíl maximální délky do pozice i a mimimální délky do pozice i

$$\bar{x}_i = \frac{1}{i}\sum_{j=1}^{i} x_i \ .$$

$\bar{x}_1$ je průměrná délka jednotky do pozice i,

$$S_i = \left[\frac{1}{i}\sum_{j=1}^{i}(x_j - \overline{x_i})^2\right]^{1/2}$$

a vypočítejte pro každý krok R_i/S_i. Sekvence může být vyhlazena, pokud se bude počítat s hodnotami i = 10, 20, 30,... Aplikujte funkci $R_i/S_i = ai^H$ na vytvořené sekvence a interpretujte chování těchto sekvencí za použití vhodné literatury. Vypočítejte míru korelace a Hausdorffovu-Besicovitchovu dimenzi, vyvoďte nějaké závěry týkající se textových sekvencí. Porovnejte více jazyků.

Vezměte další kvantifikovatelné vlastnosti texu a analyzujte jejich sekvence stejným způsobem. Nakonec se pokuste najít základ chaotického chování lingvistických sekvencí.

Literatura

Çambel, A. B. (1993). *Applied chaos theory. A paradigm for complexity.* San Diego: Academic Press.

Feder, J. (1988). *Fractals*. New York: Plenum.

Hřebíček, L. (1997). Persistence and other aspects of sentence-length series. *Journal of Quantitative Linguistics 4(1–3)*, 103–109.

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

Hurst, H. E., Black, R. P., Simaika, Y. M. (1965). *Long term storage, an experimental study*. London: Constable.

Mandelbrot, B. (1982). *The fractal geometry of nature*. New York: Freeman.

Mandelbrot, B., Wallis, J. R. (1969a). Some long-run properties of geophysical records. *Water Resources Research 5(2)*, 321–340.

Mandelbrot, B. Wallis, J. R. (1969b). Robustness of the rescaled range R/S in the measurement on noncyclic long run statistical dependence. *Water Resources Research 5(5)*, 967–988.

4.12 KÖHLEROVY MOTIVY DÉLEK SLOV 1

Hypotéza

Motivy délek jsou jazykové jednotky, které se chovají stejně jako ostatní jazykové jednotky.

Postup

Motiv délky je sekundární jednotka definovaná jako neklesající sekvence délek primárních jednotek, např. sekvence délek slov. Pokud je např. délka slov měřena v počtu slabik, pak se věta „Motivy délky jsou lingvistické jednotky, které se chovají stejně jako jiné jednotky" skládá ze sekvence délek

3-2-1-4-3-2-1-3-2-2-2-3,

v níž jsou motivy

3, 2, 1-4, 3, 2, 1-3, 2-2-2-3.

Počet takovýchto motivů je v textu konečný, takže je možné spočítat jejich frekvenci a zkoumat jejich frekvenční distribuci. Spočítejte všechny

motivy délky v dlouhé básni a zjistěte, zda jejich ranková frekvenční distribuce je identická s rankovou frekvenční distribucí délek slov.

Pokuste se vytvořit typologii textů na základě parametrů příslušné distribuce. Pokud zkoumáte více jazyků, sledujte mezi nimi rozdíly.

Zjistěte, kolik různých sekvencí můžeme stanovit za předpokladu, že (hypoteticky) je největší délka slova L a největší sekvence délky je R. Pokud $L = R = 3$, jsou možné následující sekvence:

2, 3,

1-2, 1-3, 2-2, 2-3, 3-3,

1-1-2, 1-1-3, 1-2-2, 1-2-3, 1-3-3, 2-2-2, 2-2-3, 2-3-3, 3-3-3,

…

1-1-1-…-2, 1-1-1-…-3, ….

Vytvořte vzorec pro stanovení počtu možných motivů. Určete bohatství motivů jako poměr mezi motivy, které se vyskytly, a všemi potenciálními motivy. Stejným způsobem vymezte četnost segmentů (viz také Uhlířová 2007).

Literatura

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: Genzor, J., Bucková, M. (eds.), *Favete linguis. Studies in honour of Victor Krupa*. Bratislava: Academic Press, 142–152.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: *Data Analysis, Machine Learning and Applications*. Springer Berlin Heidelberg, 637–645.

Lua, K. T. (1990). Analysis of Chinese character stroke sequences. *Computer Processing of Chinese & Oriental Languages 6(2)*.

Uhlířová , L. (2007). Word frequency and position in sentence. *Glottometrics 14*, 1–20.

4.13 KÖHLEROVY MOTIVY DÉLEK SLOV 2

Problém

Köhlerovy motivy délky mají své specifické délky. Najděte distribuci délek motivů délek.

Postup

V předchozím problému jsou tyto délky motivů: 1, 1, 2, 1, 2, 4. Vytvořte distribuci délek Köhlerových motivů v textu a zkoumejte, zda je tato distribuce identická s distribucí délek slov? Liší se texty v této vlastnosti? Vypočítejte momenty těchto distribucí a zobrazte Ordovo schéma (viz část 8.9, „Ordovo kritérium"). Zkuste najít rozdíly mezi texty různých žánrů.

Literatura

Best, K.-H. (ed.) (1997). *The distribution of word and sentence length*. Trier: Wissenschaftlicher Verlag.

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: Genzor, J., Bucková, M. (eds.), *Favete linguis. Studies in honour of Victor Krupa*. Bratislava: Academic Press, 142–152.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: *Data Analysis, Machine Learning and Applications*. Springer Berlin Heidelberg, 637–645.

4.14 KÖHLEROVY MOTIVY DÉLEK SLOV 3

Problém

Vytvořte distribuci motivů délek slov za použití počtu morfémů jako délkových jednotek.

Postup

Spočítejte počet morfémů v analyzovaných slovech. Potom proveďte stejný postup jako v dvou předchozích problémech, přičemž berte ohled na rozdílnou definici motivů délek.

Pokud zkoumáte japonštinu, pokuste se provést všechny výpočty na morách jakožto jednotkách určujících délku motivu. Pokud zkoumaný jazyk nemá jasně oddělené hranice slov, určete je. Nezapomeňte přitom, že všechny výsledky jsou ovlivněny těmito výchozími podmínkami. Pokud se změní segmentace slov, mohou se změnit i výsledky.

Köhlerovy motivy délek jsou analogické k rytmickým jednotkám, které však mají pouze délku, ale nemají kombinatorické možnosti. Srov. rytmické jednotky jako např. 1, 1-0, 1-0-0, 1-0-0-0,…(1 označuje přízvučnou slabiku, 0 nepřízvučnou).

Literatura

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: Genzor, J., Bucková, M. (eds.), *Favete linguis. Studies in honour of Victor Krupa*. Bratislava: Academic Press, 142–152.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: *Data Analysis, Machine Learning and Applications*. Springer Berlin Heidelberg, 637–645.

4.15 KÖHLEROVY MOTIVY DÉLEK VĚT

Problém

Délky vět v textu můžeme převést na sekvence čísel. Ověřte, zda všechny předchozí problémy týkající se motivů mohou být aplikovány na věty.

Postup

Vypočítejte Hurstův exponent, Minkowského klobásu a Ljapunovův koeficient pro délku vět a sledujte rozdíly těchto indikátorů mezi různými autory, žánry a jazyky. Vypočítejte autokorelaci a porovnejte texty.

Literatura

Schils, E., Haan, P. de (1993). Characteristics of sentence length in running text. *Literary and Linguistic Computing 8(1)*, 20–26.

4.16 LORENZOVA KŘIVKA

Problém

Protřednictvím Lorenzovy křivky charakterizujte rankovou frekvenční distribuci slov.

Postup

Vytvořte rankovou frekvenční distribuci slov z krátkého textu. Vyneste odpovídající Lorenzovu křivku. Způsob vytvoření Lorenzovy křivky najdete na mnoha internetových stránkách. Pokuste se použít nějakou vlastnost této křivky k určení slovního bohatství. Proveďte stejný postup u Giniho koeficientu.

Literatura

Popescu, I.-I., Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics 13*, 23–46.

Popescu, I.-I., Vidya, M. N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008). *Word frequency studies*. Berlin, New York: de Gruyter.

4.17 LJAPUNOVŮV KOEFICIENT

Problém

Ljapunovův koeficient byl uveden do lingvistiky Hřebíčkem (1997, 2000), ale jeho význam není zcela jasný. Proveďte experimenty s texty různých autorů, žánrů a jazyků, použijte při tom různé jazykové jednotky.

Postup

Převeďte text na sekvence hodnot nějakých proměnných, např. délek slov, délek vět, polysémie, délek rytmických jednotek atd. Nechť jsou jednotlivé hodnoty x_i. Vypočítejte Ljapunovův koeficient

$$\lambda = \frac{1}{k} \sum_i \ln |x_i - x_{i+1}| \ ,$$

kde k je počet rozdílů. Nulové rozdíly by měly být ze sumy odstraněny (kvůli logaritmu). Tento koeficient je používán při výzkumu vlastností chaosu.

Pokuste se interpretovat tento koeficient za pomoci odborné literatury. Pokuste se odvodit jeho variance, využijte přitom toho, že $V(x) = \sigma^2$.

Literatura

Çambel, A. B. (1993). *Applied chaos theory. A paradigm for complexity*. San Diego: Academic Press.

Falconer, K. (1990). *Fractal geometry. Mathematical foundations and applications*. Chichester: Wiley.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute.

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

Schroeder, M. (1991). *Fractals, chaos, power laws*. New York: Freeman.

Schuster, H. G. (1995). *Deterministic chaos. An introduction*. Weinheim: VCH.

4.18 MINKOWSKÉHO KLOBÁSA

Hypotéza

„Způsob, jakým jsou jazykové konstrukty řazeny do pozičních řad, odpovídá takovému řádu nárůstu α…, jehož výsledkem je vztah podobný vztahu mezi konstrukty a jejich konstituenty (tj. Menzerathovu zákonu), pokud jsou definovány podle spojitosti a radiusu odpovídající Minkowského posloupnosti." (Hřebíček 2000: 76).

Postup

Hypotéza říká, že „Minkowského klobása" sekvenčních vlastností textu se řídí mocninným zákonem. Nejdříve si přečtěte čtvrtou kapitolu v Hřebíček (2000: 66–76), kde najdete aplikaci na posloupnosti délek jednotek.

Přepište text do formy číselné sekvence délek slov (počítáno v počtu fonémů, slabik, morfémů atd.) nebo délek vět (ty mohou být měřeny různými způsoby). Potom vypočítejte radius ε a zkontrolujte, zda je vzdálenost d mezi sousední sekvencemi větší než 2ε. Vzdálenost mezi sousedními elementy x_i a x_{i-1} je definována jako

$$d_i = [(x_i - x_{i-1})^2 + 1]^{1/2}.$$

Pokud je $d_i > 2\varepsilon$, pak máme *zlom*, jinak máme *spojitost*. Sčítejte všechna d_i spojitostí. Potom pokračujte zvyšováním ε na deset různých hodnot a pro každou zvlášť sečtěte spojitosti. Následně stanovte empirickou relaci $y = \log(spojitost)/\log(\varepsilon)$ a aplikujte na data funkci $y = a\varepsilon^{-b}$. Sledujte hodnoty parametrů a a b v různých textech, žánrech nebo jazycích. Vyvoďte textologické, typologické a obecněLingvistické závěry. Najděte vlastnosti, které mají stejné parametry v různých jazycích. Zkuste vypočítat Minkowski-Bouligandovu dimenzi pro vaši sekvenci. Sledujte, jak se parametry mocninné funkce liší u různých jednotek.

Literatura

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

Schroeder, M. (1991). *Fractals, chaos, power laws*. New York: Freeman.

Tricot, C. (1995). *Curves and fractal dimension*. New York: Springer.

4.19 N-GRAMY A MOTIVY DÉLEK

Hypotéza

V textu existují motivy délek, jejichž vlastnosti jsou závislé na stylu, žánru a autorovi. Dokažte, že jejich n-gramy vykazují určité pravidelnosti.

Postup

Definujte nějakou jednotku a nějakou vlastnost, které mohou být jednoznačně měřeny v rámci textu. Takovou jednotkou může být například slovo a jednou z jeho vlastností délka. Potom převeďte text na délky jednotlivých jednotek. Tím získáte sekvenci čísel reprezentující časové řady – Markovův řetězec.

Nejdříve spočítejte frekvence jednotlivých délek a vytvořte empirickou distribuci. Potom zkoumejte bigramy, vytvořte jejich distribuci a sledujte rozdíl mezi touto distribucí a distribucí délek. Nakonec zkoumejte postupně trigramy až dekagramy.

Zaměřte se na následující otázky:

(1) Existují nějaké n-gramy, které se vyskytují častěji, než se očekává?

(2) Kolik typů n-gramů se v textu nerealizuje? Pokuste se formálně vyjádřit míru vynechaných n-gramů. Které typy n-gramů mizí v souvislosti s růstem délky (zvyšující se n)? Liší se inventáře a frekvence n-gramů u různých autorů, stylů, žánrů, historických období, jazyků atd.?

Definujte další vlastnosti textu a zkoumejte jejich n-gramy. Najděte další problémy související s Köhlerovými motivy.

Literatura

Brainerd, B. (1976). On the Markov nature of text. *Linguistics 176*, 5–30.

Damashek, M. (1995). Gauging similarity with N-grams: language--independent categorization of text. *Science 267*, 843–848.

Egghe, L. (1999). On the law of Zipf-Mandelbrot for multi-word phrases. *Journal of the American Society for Information Science 50(3)*, 843–848.

Egghe, L. (2000). The distribution of N-grams. *Scientometrics 47(2)*, 237–252.

Kjell, B. (1994). Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing 9(2)*, 119–124.

Köhler, R. (1983). Markov-Ketten und Autokorrelation in der Sprach- und Textanalyse. *Glottometrika 5*, 134–167.

Lua, K. T. (1995). *A minimum entropy approach for Chinese text compression.* Dostupné z: *http://www.iscs.nus.sg/~luakt*

Mayzner, M. S., Tresselt, M. E., Wolin, B. R. (1965). Tables of tetragram frequency counts for various word-length and letter-position combinations. *Psychonomic monograph supplements 1(4)*, 79–143.

Robertson, A. M., Willet, P. (1998). Applications of N-grams in textual information systems. *Journal of Documentation 54(1)*, 48–69.

Runquist, W. N. (1968). Rated similarity of high m CVC trigrams and words and low m CCC trigrams. *Journal of Verbal Learning and Verbal Behavior 7*, 967–968.

Schönpflug, W. (1969). n-Gramm-Häufigkeiten in der deutschen Sprache. I. Monogramme und Digramme. *Zeitschrift für experimentelle und angewandte Psychologie XVI*, 157–183.

Siméonoff, E. (1965). On the distributions of the „costs" of combinations of K letters in a written language. *Statistical Methods in Linguistics 4*, 45–50.

Suen, C. Y. (1979). n-Gram statistics for natural language understanding and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1/2*, 164–172.

Willett, P. (1979). Document retrieval experiments using indexing vocabulary of varying size II. Hashing, truncation, digram and trigram encoding of index terms. *Journal of Documentation 35(4)*, 296–305.

Yannakoudakis, E. J., Tsomokos, I., Hutton, P. J. (1990). n-Grams and their implication to natural language understanding. *Pattern Recognition 23(5)*, 509–528.

4.20 NOMINÁLNÍ STYL

Problém

Nominální styl je někdy dáván do kontrastu k slovesnému stylu. Pokuste se vyjádřit rozdíl mezi oběma styly kvantitativně.

Postup

Spočítejte počet substantiv (N) a sloves (V) v textu. Ostatní slova nejsou důležitá. Proveďte následující test

$$X^2 = \frac{(N-V)^2}{N+V}$$

nebo případně

$$z = \left(\frac{2N}{R} - 1\right)\sqrt{R}\,,$$

kde $R = N + V$. Oba testy jsou ekvivalentní. X^2 je chí-kvadrát test s 1 stupněm volnosti a kritickou hodnotou 3,84; z je normální test, $z^2 = X^2$ a kritická hodnota je 1,96. Interpretujte výsledky testu.

Porovnejte styl lyrické a epické poezie, styl vědecký a publicistický. Popište svá zjištění.

Literatura

Ziegler, A., Best, K.-H., Altmann, G. (2002). Nominalstil. *ETC – Empirical Text and Culture Research* 2, 72–85.

4.21 FONETICKÁ AGREGACE

Problém

Podle Skinnerovy hypotézy existuje, v rámci krátké vzdálenosti, zvyšující se pravděpodobnost, že se již jednou použitá jednotka bude opakovat. Skinner vysvětlil tento jev na základě předpokladu zvyšující se aktivity zapojených neuronů. V důsledu toho dochází k tomu, že ve spontánním mluveném projevu jsou si ty bloky textu, jako jsou věty nebo verše, jež jsou umístěny blízko sebe, foneticky podobnější než ty, které leží dále od sebe. Tento jev může být prokázán zejména u spontánně vyprávěné lidové poezie. Proveďte následující:

(1) Testujte, zda hypotéza platí v díle Goetheho, Shakespeara nebo Ovidia.

(2) Zjistěte, zda zmenšující se fonetická podobnost jednotek související z jejich zvyšující se vzdáleností může být považována za znak spontánnosti.

Postup

Proveďte fonetickou transkripci zvolené básně. Vytvořte způsob měření fonetické podobnosti veršů ve vzdálenostech $x = 1, 2, 3,\ldots$ Určete, zda existuje klesající tendence a navrhněte vzorec vyjadřující tento pokles,

zřejmě $y = ax^{-b}$. Pokud bude tendence skutečně nalezena, můžeme tvrdit, že spontánnost souvisí s fonetickou podobností?

Porovnejte lidovou poezii s moderní poezií. Analyzujte texty jednotlivých postav v dramatu. Jsou si více podobné jednotlivé pasáže různých postav, nebo pasáže jedné postavy? Pokud ano, pak může mít drama velmi komplexní fonetickou strukturu.

Literatura

Altmann, G. (1968). Some phonic features of Malay shaer. *Asian and African Studies 4*, 9–16.

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Skinner, B. F. (1939). The alliteration in Shakespeare´s sonnets: A study in literary behaviour. *Psychological Record 3*, 186–192.

Skinner, B. F. (1941). A quantitative estimate of certain types of sound-patterning in poetry. *The American Journal of Psychology 54*, 64–79.

4.22 ANALÝZA POLYLOGU

Problém

Frekvence i sekvence mluvních aktů v divadelní hře vytváří mnoho problémů, které mohou být řešeny v rámci nějakého projektu.

Postup

(1) Vypočítejte rankovou frekvenční distribuci a frekvenční spektrum pro každou postavu zvlášť. Seřaďte postavy podle počtu slov a porovnejte parametr jednotlivých distribucí s jejich pořadím.

(2) Vypočítejte distribuci délek vět každé postavy a porovnejte průměr těchto délek s důležitostí postavy.

(3) Klasifikujte mluvní akty a vypočítejte pro každou postavu vektor, jehož prvky budou tvořit proporce různých mluvních aktů. Použijte měření vzdálenosti (nebo blízkosti) pro klasifikaci postav.

(4) Vypočítejte matici přechodových pravděpodobností posloupnosti mluvčích a zkoumejte jejich vlastnosti. Vyvoďte závěry o interakci postav. Vytvořte vážený graf interakcí.

(5) Vytvořte sekvenci mluvních aktů tak, že je označíte písmeny a zkoumejte vlastnosti této sekvence. Tvoří se nějaké nepřerušené řady posloupností stejných písmen?

(6) Škálujte mluvní akty v určité dimenzi (např. postoj) a sledujte chování této sekvence. Nepoužívejte Fourierovy řady, ale pokuste se zkoumat její fraktální dimenzi.

(7) Zkoumejte (na základ formálních kritérií) postoj každé postavy k ostatním postavám na základě mluvních aktů.

(8) Pokud mluvní akty nějakým způsobem změříte, je možné definovat motivy mluvních aktů. Zkoumejte jejich distribuci a sekvence.

(9) Porovnejte divadelní hry různých žánrů, např. tragédie s komediemi.

(10) Porovnejte jednotlivé hry jednoho autora, zkuste najít charekteristiky jeho historického vývoje. Použijte různé charakteristiky textu.

(11) Proveďte analogické výpočty aplikované na druhy vět. Tyto musí být předem přesně definovány.

(12) Pokud je to možné, škálujte mluvní akty sémanticky pomocí Osgoodova sémantického diferenciálu s vhodnými dimenzemi (Osgood et al. 1957).

Literatura

Osgood, C. E., Suci, G. J., Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: Univ. Illinois Press.

Snider, J. G., Osgood, C. E. (1969). *Semantic Differential Technique: A Sourcebook*. Chicago: Aldine.

Indiana University. [online]. Dostupné z: http://www.indiana.edu/~socpsy/papers/AttMeasure/attitude.htm

4.23 POPESCŮV INDEX SLOVNÍHO BOHATSTVÍ

Problém

Čím více slov s nízkou frekvencí je v textu použito, tím je větší jeho slovní bohatství. Jedním ze způsobů výpočtu slovního bohatství textu je Popescův index R_1. Pokuste se pomocí tohoto indexu charakterizovat různé texty a autory.

Postup

Protože autosémantika, která přispívají ke slovnímu bohatství textu, mají obvykle vyšší rank než h (viz část 7.6, „Popescův typologický a-indikátor"), Popescu et al. 2008 navrhl následující index:

$$R_1 = 1 - \left(F(h) - \frac{h^2}{2N} \right) ,$$

kde $F(h)$ je kumulativní relativní frekvence slov, jejichž pořadí je nižší nebo rovno h, h značí h-bod, N je délka textu (měřena ve slovních formách). Zpracujte různé texty, seřaďte slova podle jejich frekvence, vypočítejte h a $F(h)$ a výše uvedený index.

Pokud chcete porovnávat texty na základě rozdílů ve slovním bohatství, testujte rozdíly mezi dvěma indexy R_1 pomocí následující rovnice:

$$z = \frac{R_{1,1} - R_{1,2}}{\sqrt{Var(R_{1,1}) + Var(R_{1,2})}} ,$$

kde

$$Var(R_{1,i}) = F(h_i)[1\text{-}F(h_i)]/N \; (i = 1,2).$$

Pokuste se pomocí slovního bohatství charakterizovat texty, autory a žánry. Další indexy pro výpočet slovního bohatství naleznete v uvedené literatuře.

Literatura

Popescu, I.-I., Vidya, M. N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008). *Word frequency studies*. Berlin, New York: de Gruyter.

4.24 INDEXY

Problém

Při analýzách stylu bývají používány nejrůznější indexy. Nejznámnější je Busemannův index vyjadřující poměr sloves a adjektiv: (*počet sloves*)/(*počet adjektiv*). Navrhněte další různé indexy, normalizujte je, odvoďte variance a vytvořte asymptotický test. Daný index interpretujte.

Postup

Nejdříve navržený index normalizujte, tj. jeho výsledná hodnota musí ležet v intervalu <0, 1>. Bussemanův index není normalizovaný, proto není ve své původní podobě interpretovatelný. Nejjednodušším způsobem vytvoření indexu je vyjádření proporce, v níž je variance dána automaticky, díky čemuž může být jednoduše vytvořen test.

Aplikujte daný index na různé texty a porovnejte je pomocí vámi vytvořeného testu. Následně interpretujte výsledky. Pokuste se najít

signifikantní rozdíly mezi texty nebo žánry a zjistěte, zda vaše metoda může být použita ve stylometrii.

Literatura

Altmann, G. (1978). Zur Verwendung der Quotiente in der Textanalyse. *Glottometrika 1*, 91–106.

Tuldava, J. (2005). Stylistic author identification. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, 369–387.

4.25 RYTMICKÉ JEDNOTKY

Problém

Sekvence přízvučných a nepřízvučných slabik vytvářejí v prozaickém textu určitý druh rytmu, který vykazuje různé vlastnosti. Najděte některé vlastnosti a pravidelnosti jejich chování.

Postup

Definujte rytmickou jednotku jako posloupnost přízvučné slabiky, za níž následují slabiky nepřízvučné (v poezii může být definována jinak). Přízvučná slabika může být označena jako 1, nepřízvučná jako 0. Získáme tak jednotky 1, 10, 100, 1 000,… Délky těchto jednotek mohou být definovány počtem slabik, které je tvoří. Přepište text do formy sekvence délek těchto jednotek. Například sekvence 101001010000110 bude přepsána na 2, 3, 2, 5, 1, 2. Potom zkoumejte:

(1) Jaká je distribuce délek? Existuje jedno obecné rozdělení platné pro všechny prozaické texty, nebo jsou mezi texty rozdíly? Testujte Hyperpoissonovo rozdělení.

(2) Vytvořte tabulku přechodových pravděpodobností mezi délkami a zkoumejte sekvence jako Markovovy řetězce. Určete řád řetězce. Vypočítejte limitní stav pravděpodobnosti vektoru matice přechodových pravděpodobností.

(3) Považujte vektor v 2. bodu za charakteristiku textu a porovnejte různé texty na základě Euklidovské vzdálenosti.

(4) Spočítejte frekvence bigramů, trigramů atd., vytvořte jejich distribuce a vypočítejte jejich entropii. Pokuste se modelovat průběh entropie od monogramů k n-gramům (vzhledem k délce textu). Najděte nejmenší n (přímo nebo extrapolací), pro něž entropie dosahuje svého maxima (tj. kde se všechny n-gramy vyskytují právě jednou).

(5) Zkoumejte autokorelaci symbolů (0, 1) a zvlášť autokorelaci délek až do $k = 20$. Vytvořte graf.

(6) Zkoumejte vzdálenosti mezi jednotkami stejných délek. Jsou náhodné, nebo se řídí nějakou tendencí? Míru náhodnosti stanovte prostřednictvím Zörnigova rozdělení vzdáleností.

Literatura

Best, K.-H. (2002). The distribution of rhythmic units in German short prose. *Glottometrics 3*, 136–142.

Best, K.-H. (2005). Längen rhythmischer Einheiten. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, 208–214.

Eom, J. (2006). *Rhythmus im Akzent. Zur Modellierung der Akzentverteilung als einer Grundlage des Sprachrhythmus im Russischen*. München: Sagner.

Marbe, K. (1904). *Über den Rhythmus der Prosa*. Giessen: J. Ricker´sche Verlagsbuchhandlung.

Zörnig, P. (1984a). The distribution of distances between like elements in a sequence I. *Glottometrika 6*, 1–15.

Zörnig, P. (1984b). The distribution of distances between like elements in a sequence II. *Glottometrika 7,* 1–14.

Zörnig, P. (1987). A theory of distances between like elements in a sequence. *Glottometrika 8,* 1–22.

4.26 OBTÍŽNOST TEXTU

Problém

Zamyslete se nad problémem měření obtížnosti textu.

Postup

Po nastudování dostatečného množství literatury shromážděte všechny vlastnosti ovlivňující obtížnost textu a krok za krokem je analyzujte. Pokuste se vysvětlit vliv těchto vlastností na obtížnost (čitelnost) textu. Analýzy některých vlastností můžete najít v Kukemelk, Mikk (1993).

Shromážděte všechny vzorce, které se používají pro výpočet obtížnosti (srozumitelnosti) textu, a zkoumejte jejich slabé stránky. Pokuste se najít závislosti mezi těmito vlastnostmi a odstraňte ty, které se jeví jako redundantní. Pokud budete navrhovat vlastní způsob výpočtu, vytvořte vzorec tak, aby mohly být porovnávány dva texty.

Literatura

Kukemelk, H., Mikk, J. (1993). The prognosticating effectivity of learning a text in physics. *Glottometrika 14*, 82–103.

4.27 TEMATICKÁ KONCENTRACE

Problém

Tematická koncentrace se vztahuje k určité množině slov v textu. Vypočítejte tematickou koncentraci poetického a vědeckého textu.

Postup

Spočítejte frekvence slov v textu. Lepší výsledky získáte při analýze lemmatizovaného textu. Potom vytvořte rankovou frekvenční distribuci a vypočítejte *h*-bod (viz část 7.6, „Popescův typologický *a*-indikátor"). Zaměřte se pouze na autosémantická slova nad *h*-bodem (tj. $r' \leq h$). V próze se v této pozici objevují i vlastní jména, u kterých musíte rozhodnout, zda je řadit mezi tematická slova. Popescův index tematické koncentrace je definován takto:

$$TC = 2 \sum_{r'=1}^{T} \frac{(h - r')f(r')}{h(h-1)f(1)} \ ,$$

kde h = *h*-bod,

r' = rank slova nad *h*-bodem,

$f(r')$ = frekvence tematického slova, jehož rank je r',

$f(1)$ = četnost nejfrekventovanějšího slova,

T = počet ranků tematických slov nad *h*-bodem.

Někdy může být tematické slovo reprezentováno prostřednictvím více výrazů, jejichž význam je totožný, např. *Julie*, *mladá dívka,* atd. Můžete započítávat i frekvence všech výrazů, které reprezentují takové tematické slovo. Výsledkem bude jiná charakteristika této vlastnosti a jiná hodnota tematické koncentrace.

Analyzujte mnoho textů a zkuste najít rozdíly tematické koncentrace v jednotlivých žánrech. Potom seřaďte žánry podle tematické koncentrace.

Literatura

Popescu, I.-I., Vidya, M. N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008). *Word frequency studies.* Berlin, New York: de Gruyter.

4.28 TOKENY A LJAPUNOVŮV KOEFICIENT

Problém

Většinu slov v textu může autor nahradit určitým množstvím alternativ, které nemění význam daného slova, ale přidávají mu jisté významové nuance. Množina těchto alternativ a dané slovo tvoří token M_k o velikosti $|M_k|$, kde k je pozice v textu. K textu lze tedy přistupovat jako k sekvenci tokenů. Nás zajímá pouze *sekvence velikosti tokenů*, která zobrazuje *autorský informační obsah* (možnost volby) na dané pozici. Vypočítejte Ljapunovův koeficient pro takový druh sekvence.

Postup

Dosud byl tímto způsobem analyzován pouze jediný text (Andersen, Altmann 2006). Na základě hodnot uvedených v tabulce na s. 109–115 (Velikost tokenu $|M_k|$) v příloze citovaného článku vypočítejte Ljapunovův koeficient pro tuto sekvenci.

Pokud je vámi zkoumaný jazyk vaším rodným jazykem, pokuste se tímto způsobem analyzovat několik krátkých textů. Vyvoďte závěry o jednotlivých textech a žánrech, zkuste prozkoumat *autorský* informační tok obecně. Sledujte také další charakteristiky sekvence tohoto druhu.

Literatura

Andersen, S., Altmann, G. (2006). Information content of words in texts. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues*. Dordrecht: Springer, 91–115.

4.29 VZTAH TYPŮ A TOKENŮ

Problém

Vypočítejte fraktální dimenzi Köhler-Galleovy sekvence typů a tokenů.

Postup

Köhler-Galleův poměr typů a tokenů (TTR) je vyjádřen vzorcem:

$$TTR_x = \frac{t_x + T - \frac{xT}{N}}{N},$$

kde x = pozice v textu (počet tokenů do pozice x),

t_x = počet typů do pozice x,

T = počet typů v celém textu,

N = délka textu (počet tokenů v celém textu).

Na základě nějakého textu vypočítejte tuto sekvenci (použijte výše uvedený vzorec). Potom vytvořte graf sekvence, ten bude mít fraktální charakter. Vypočítejte různé druhy fraktálních dimenzí.

Proveďte výpočet u různých textů a různých jazyků. Zkoumejte jejich podobnosti a rozdíly.

Literatura

Falconer, K. J. (1990). *Fractal geometry. Mathematical foundations and applications*. Chichester: Wiley.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute.

Köhler, R., Galle, M. (1993). Dynamic aspects of text characteristics. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative text analysis*. Trier: Wissenschaftlicher Verlag, 46–53.

Schroeder, M. (1991). *Fractals, chaos, power laws. Minutes from an infinite paradise*. New York: Freeman.

Tricot, C. (1993). *Curves and fractal dimensions*. New York: Springer.

Wimmer, G. (2005). The type-token relation. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 361–368.

4.30 SLOVESNÝ PROFIL

Problém

Jednou z možných charakteristik textu je jeho slovesný profil, tj. komplexní obraz reprezentující jeho „slovesné chování".

Postup

Vytvořte slovesný profil textu. Levickij and Lučak (2005) navrhli následující sémantické třídy sloves:

(1) Slovesa směny (*kupovat, prodávat, platit, měnit, obchodovat*).

(2) Slovesa míry (*účtovat, pokutovat, stát [o ceně], odhadnout, pokutovat, měřit, ocenit,vážit*).

(3) Slovesa změny vlastníka (*dát, vzít, dostat, půjčit, ukrást, vrátit*).

(4) Slovesa změny pozice (*klesat, padat, hodit, klouzat [se], letět, plavat, posunout, otočit, přesunout*).

(5) Změna fyzikálního stavu (*tavit, rudnout, změknout, mrazit, sušit, tvrdnout, zlomit*).

(6) Okolnostní slovesa (*začít, skončit, opakovat, zahájit, pokračovat, ukončit, zastavit, dokončit, spustit, zastavit, udržovat*).

(7) Slovesa následku (*ustřihnout, bodnout, prorazit, rozbít, propíchnout, kousnout, zastřelit, zabít*).

(8) Slovesa směru pohybu (*vstupit, přijít, jít, dorazit, klesat, stoupat, zvednout, snížit, vystoupit, povstat, odjet, vrátit se*).

(9) Slovesa existence (*existovat, žít, pobývat, bydlet, objevit se, zmizet, zůstat*).

(10) Slovesa požití (*ukousnout, vypít, sníst, hltat, polykat, cucat*).

(11) Slovesa mentálních procesů (*hádat, vědět, učit se, pamatovat, studovat, myslet*).

(12) Slovesa shromažďování/rozptýlení (*rozházet, rozprášit, hromadit, zabalit*).

(13) Slovesa způsobu pohybu (*skákat, tančit, plavit se, chodit, procházet se*).

(14) Slovesa vlastnictví (*náležet, patřit, mít, vlastnit*).

(15) Slovesa percepce a komunikace (*ptát se, komunikovat, slyšet, poslouchat, dívat se, vidět, cítit, mluvit, povídat, říkat, sledovat*).

(16) Poziční slovesa (*zůstat, stát*).

(17) Slovesa odstranění (*eliminovat, odstranit, vyprázdnit, vydrhnout, zamést, sloupnout, vyloupnout*).

(18) Slovesa orientace (*cílit, čelit, orietovat se, ukazovat*).

(19) Slovesa psychického stavu (*bavit, otravovat, strašit, užít si, nenávidět, mít rád, líbit se*).

(20) Slovesa hluku (*řvát, tlachat, pískat, štěkat, dunět, pištět*).

Pokud bychom použili všechny třídy, vektor by měl 20 prvků. Pokuste se různými způsoby sloučit některé třídy, abyste získali menší vektor. Použijte jiné klasifikace sloves. Pokuste se použít různé škálování těchto tříd, např. vývojové pořadí, nebo od existence přes pohyb, jezení, hmatání, uchopování, prohlížení atd. až k mentálním procesům atd. Potom vytvořte relevantní vektory. Pro uvedenou výše klasifikci bude

$V = \{e1, e2, \ldots e20\}$.

Potom vyberte různé texty a spočítejte (a) počet typů (sloves) náležejících ke každé z těchto 20 tříd, (b) počet tokenů náležejících ke každé z těchto 20 tříd. V případě potřeby můžete hodnoty normalizovat. Vektor změňte podle vašeho škálování. Porovnejte texty a dokažte, že slovesný profil se u různých žánrů liší.

V dalším kroku nepovažujte hranice jednotlivých druhů za jednoznačné. Označte u každého slovesa míru, se kterou náleží do nějaké třídy. Pokuste se pracovat s fuzzy množinami.

Literatura

Halliday, M. A. K. (1994). *An introduction to functional grammar*. London: Arnold.

Levickij, V. V., Lučak, M. (2005). Category of tense and verb semantics in the English language. *Journal of Quantitative Linguistics 12(2–3)*, 212–238.

Levin, B. (1999). *English verb classes and alternations*. Chicago: University of Chicago Press.

Scheibman, J. (2001). Local pattern of subjectivity in person and verb type in American English conversation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 60–89.

4.31 BOHATSTVÍ SLOVNÍKU A REFERENCE

Hypotéza

Hřebíček (1985) vycházel ze dvou předpokladů. (1) Čím větší je slovní bohatství textu, tím menší je počet referencí, (2) čím více vět je v textu, tím více je tam referencí. Hřebíček odvodil vzorec:

$$r = csn^b,$$

kde r = počet referencí, s = počet vět v textu, n = délka textu (počet tokenů, např. slovních tvarů), c a b jsou parametry.

Postup

Definujte přesně pojem reference, potom analyzujte několik textů. Existují rozdíly v parametru b pro (a) jednotlivé autory, (b) různé žánry, (c) jednotlivé jazyky? Parametry b a c v Hřebíčkově vzorci mohou být odhadnuty z dat klasickou metodou nebo pomocí algoritmů při iterativní optimalizaci. Porovnejte vaše výsledky s Hřebíčkovými. Vytvořte nový vzorec a teoreticky jej zdůvodněte.

Literatura

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Hřebíček, L. (1985). Text as a unit and co-references. In: Ballmer, T. T. (ed.), *Linguistic dynamics*. Berlin, New York: de Gruyter, 190–198.

4.32 FREKVENCE SLOV 1

Problém

S rankovou frekvenční distribucí slov v textu souvisí mnoho problémů, které ještě nejsou zcela vyřešeny. Pokuste se některé z nich zkoumat.

Postup

(1) Vytvořte rankovou frekvenční distribuci slov (lemmat, nikoliv slovních tvarů) v textu.

(2) Převeďte ji na *kumulativní* rankovou frekvenční distribuci (empirickou distribuční funkci). Najděte empirickou spojitou funkci, kterou je možné adekvátně aplikovat na data. Může to být i polynom.

(3) Vypočítejte *délku křivky* prostřednictvím standardních vzorců pro analýzu tohoto typu.

(4) Zkuste najít odpovědět na otázku: *Souvisí nějak tato délka křivky s bohatsvím slovníku*? Pokud ano, vysvětlete tuto souvislost, tj. určete, zda delší křivka koreluje s větším slovním bohatstvím. Zkoumejte mnoho textů (krátkých i dlouhých).

(5) Pokuste se aplikovat všechna dostupná diskrétní rozdělení na vaše data, neomezujte se jen na Zipf-Mandelbrotovu teorii.

Literatura

Baayen, R. H. (2005). Word frequency distributions. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 397–409.

4.33 FREKVENCE SLOV 2

Problém

Prostudujte historii zkoumání rankové frekvenční distribuce slov.

Postup

Projděte všechnu dostupnou literaturu. Omezte se na vzorce popisující rankové frekvenční distribuce. Sledujte rozdíly v pojetí slova (slovní tvar nebo lemma), metody výběru vzorku (náhodný výběrový soubor, kompletní text, homogenní texty atd.) a třídy slov. Začněte Estoupem (1916) a nevynechte ruské práce. Množství literatury týkající se tohoto tématu je obrovské, zde uvádíme jen seznam přehledových prací.

Literatura

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Chitashvili, R. J., Baayen, R. H. (1993). Word frequency distributions. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative text analysis*. Trier: Wissenschaftlicher Verlag, 54–135.

Orlov, J. K., Boroda, M. G., Nadarejšvili, I. Š. (1982). *Text, Sprache, Kunst. Quantitative Analysen*. Bochum: Brockmeyer.

Popescu, I.-I., Vidya, M. N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008). *Word frequency studies*. Berlin: de Gruyter.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin: de Gruyter, 791–807.

4.34 FREKVENCE SLOV 3

Problém

Obvykle je ranková frekvenční distribuce modelována Zipfovým (zeta) rozdělením. Dokažte, že Popescova-Altmannova-Köhlerova (PAK) křivka je vhodnějším modelem.

Postup

Použijte jakoukoliv rankovou frekvenční distribuci z textu. Pomocí vhodného programu pro aplikaci teoretických rozdělení na data (např. NLREG, TableCurve, Origin atd.) aplikujte mocninnou funkci $y = Cx^{-a}$ na data. Sledujte hodnotu determinačního koeficientu R^2. Potom aplikujte funkci

$$y = 1 + a^* \exp(-bx) + c^*\exp(-dx)$$

a porovnejte hodnoty determinačních koeficientů u obou modelů. Sledujte grafy obou funkcí. Mocninná funkce konverguje k nule, zatímco křivka druhého modelu konverguje k 1, a lépe tak zachycuje hapax legomena. Pokuste se modifikovat mocninnou funkci prostřednictvím $y = 1 + Cx^{-a}$ a porovnejte determinační koeficienty. Pracujte pouze s jedním komponentem PAK a porovnejte výsledky.

Literatura

Popescu, I.-I., Altmann, G., Köhler, G. (2008). *Zipf´s law – another view. Quality & Quantity, 44(4)*, 713–731.

5 Frekvence a délka

5.1 DISTRIBUCE DÉLKY SLOV 1

Problém

Meyer (1997, 1999) ve své studii o slovní délce v jazyce inuktitut objevil nové rozdělení délek (konvoluce Poissonova a Thomasova rozdělení). Analyzujte toto rozdělení a zkuste ho aplikovat na jiné jazyky.

Postup

Najděte první momenty rozdělení za použití pravděpodobností vytvářející funkce. Potom se pokuste najít odhady těchto dvou parametrů za pomoci momentů nebo frekvenčních tříd. Následně testujte toto rozdělení na empirických distribucích délek, které najdete v příslušné literatuře. Zkoumejte různé jazyky. Pokud bude rozdělení adekvátní, popište obecné rysy těchto jazyků.

Literatura

Best, K. H. (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.

Meyer, P. (1997). Word length distribution in Inuktitut narratives: empirical and theoretical findings. *Journal of Quantitative Linguitics 4,* 143–155.

Meyer, P. (1999). Relating word length to morphemic structure: a morphologically motivated class of discrete probability distributions. *Journal ov Quantitative Linguistics 6(1),* 66–69.

5.2 DISTRIBUCE DÉLKY SLOV 2

Problém

Najděte vhodné rozdělení délek slov v několika textech v jazyce, který ještě nebyl takto prozkoumán. Délku slov definujte počtem slabik. Pouze v případě, že analyzujete monosylabický jazyk, použijte fonémy jako základní jednotku pro výpočet délky slova.

Postup

Nastudujte příslušnou literaturu. Postupujte induktivně, tj. aplikujte různá rozdělení a vyberte ta, která budou vhodná pro všechny texty. Zkoumejte parametry jednotlivých distribucí a zkuste nalézt trend nebo rozdíly mezi empirickými distribucemi.

Literatura

Best, K.-H. (ed.) (1997). *The distribution of word and sentence length*. Trier: Wissenschaftlicher Verlag.

Best, K. H. (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.

Wimmer, G., Altmann, G. (1996). The theory of word length: some results and generalizations. *Glottometrika 15*, 166–180.

Wimmer, G., Köhler, R., Grotjahn, R., Altmann, G. (1994). Towards a theory of word length distributions. *Journal of Quantitative Linguistics 1*, 98–106.

Bibliographische Übersicht zum Göttinger Projekt zur Quantitativen Linguistik. [online]. Dostupné z: http://wwwuser.gwdg.de/~kbest/litlist.htm

5.3 DISTRIBUCE DÉLKY SLOV A ORDOVO KRITÉRIUM

Hypotéza

V Ordově schématu <*I*, *S*> jsou všechny distribuce délek slov textů jednoho autora umístěny na přímce.

Postup

Použijte výsledky z předchozí části 5.2, „Distribuce délky slov 2", a vypočítejte pro každý text funkci Orda.

Znázorněte vypočtené hodnoty do souřadnic <*I*, *S*> a vypočítejte přímku pro každého autora. Ve slovanských jazycích může být počítáno i s nulovou délkou slov. Porovnejte vaše výsledky s literaturou.

Literatura

Best, K. H. (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.

Best, K.–H. (2003). *Quantitative Linguistik. Eine Annäherung.* (2. Auflage). Göttingen: Peust & Gutschmidt.

5.4 FREKVENCE A TENDENCE K TVOŘENÍ KOMPOZIT

Hypotéza

„Čím vyšší je frekvence slova, tím více tvoří dané slovo kompozita." (Altmann 1989).

Postup

Sestavte náhodný výběrový soubor slov stejného slovního druhu z korpusu a zaznamenejte si jejich relativní frekvence. Potom zjistěte, v kolika

kompozitech se každé slovo vyskytuje. Abyste mohli dospět k obecným závěrům, je třeba analyzovat několik jazyků.

Seřaďte slova podle jejich frekvence a vytvořte relaci <*frekvence slova, počet kompozit*>. Výsledkem může být monotónní stoupající funkce. Pokuste se nalézt empirický vzorec a odvoďte jej z proporce.

Literatura

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10*, 46–70.

Andrukovič, P. F., Korolev, E. I. (1977). O statističeskich i leksikogrammatičeskich svojstvach slov. *Naučno-techničeskaja Informacija Serija 2*, 1–9.

Bertram, R., Schreuder, R., Baayen, R. H. (2000). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology; Learning, Memory, and Cognition 26*, 419–511.

Hay, J. (2003). *Causes and consequences of word structure*. New York: Routledge.

5.5 FREKVENCE A UŽITEČNOST PÍSMEN

Hypotéza

„Existuje vztah mezi četností písmen a jejich grafémickou užitečností." (Bernhard, Altmann 2008).

Postup

Grafémická (poziční) užitečnost písmene je měřena jako součet jeho pozic v grafémech, např. italské písmeno <g> se objevuje v grafémech <g, gl, gli, gn, gi, gg, gh, ggh> (grafém reprezentuje foném). V uvedeném příkladu se <g> objevuje osmkrát na první pozici a dvakrát na druhé pozici, tudíž

PP(g) = 8(1) + 2(2) = 12. Popište přesně vztah mezi fonémem a takto definovaným grafémem (použijte metody v knize *Analyses of script*) v daném jazyce a vypočítejte grafémickou (poziční) užitečnost každého písmene. Potom použijte korpus a získejte frekvence jednotlivých písmen. Najděte alespoň korelaci, pokud je to možné tak i funkci a teoreticky ji zdůvodněte.

Literatura

Altmann, G., Fan, F. (eds.) (2008). *Analyses of script. Properties of characters and writing systems*. Berlin, New York: de Gruyter.

Bernhard, G., Altmann, G. (2007). The phoneme-grapheme relationship in Italian. In: Altmann, G., Fan, F. (eds.), *Analyses of script. Properties of characters and writing systems*. Berlin, New York: de Gruyter, 9–19.

5.6 FREKVENCE A PŘÍZNAKOVOST/KOMPLEXITA

Hypotézy

„Nepříznakové jednotky daných kategorií jsou četnější než příznakové jednotky." (Bybee, Hopper 2001: 1).

„…existuje rovnováha mezi velikostí nebo stupněm komplexity fonému a jeho relativní četností výskytu v tom smyslu, že míra nebo stupeň komplexity fonému je inverzní vzhledem k relativní frekvenci jeho výskytu." (Zipf 1935: 49).

„…ve všech případech, kde se dá určit míra stupně komplexity fonémů, je tato míra komplexity v inverzním (ne nutně proporčním) poměru k relativní frekvenci výskytu." (Zipf 1935: 79).

„…zdá se velmi nepravděpodobné, že míra komplexity je příčinou relativní četnosti výskytu. Nicméně může být prokázáno, že opak je pravdou…" (Zipf 1935: 81).

„Přízvuk (případně míra nápadnosti) jakéhokoliv slova, slabiky nebo zvuku je inverzně proporční k relativní frekvenci daného slova, slabiky nebo zvuku mezi ostatními slovy, slabikami nebo zvuky v proudu mluvené řeči. Jakmile se nějaká jednotka začne používat častěji, její forma se stává méně přízvučnou a snadněji vyslovitelnou, a naopak. (Zipf 1929: 4).

„…termín ‚příznakovost' může být jednoduše nahrazen termínem ‚frekvence'. Frekvence je navíc skutečná empirická proměnná, zatímco příznakovost je teoretickým konstruktem." (Fenk-Oczlon 2001: 435).

„Sémantická nepříznakovost a vysoká frekvence obvykle konvergují." (Fenk-Oczlon 2001: 441).

Postup

Vytvořte seznam příznakových a nepříznakových jednotek. (a) Upřesněte hypotézu, tj. rozhodněte v jednotlivých případech, co je příznakové a co ne, (b) navrhněte metriku pro měření příznakovosti, (c) odvoďte vzorec, (d) proveďte test.

Vyberte 5 až 10 dichotomií (příznakový/nepříznakový) z různých jazykových rovin a proveďte výše uvedený postup. Potom zkuste vypočítat *stupně příznakovosti*, protože dichotomie představuje extrémní redukci informace. Třídy mohou být označeny různými stupni (srov. deklinace, konjugace). V případě nutnosti najděte pro každou jednotku různou škálovací metodu. Dobrým příkladem je Corbett, Hippisley, Brown, Marriott (2001). Porovnejte stupně příznakovosti s frekvencí jednotek, vytvořte graf a formulujte hypotézu proporcionality. Odvoďte z hypotéz křivky. Testujte je na vašich datech.

Okomentujte výrazy „skutečná empirická jednotka" a „teoretický konstrukt".

Koncept přízanakovosti najdete na s. 1, 28, 52, 54, 61, 68, 71, 82, 101, 131, 138, 140, 152–154, 185, 192, 204, 213, 215–216, 223, 226, 234, 236,

246, 292, 293, 315, 317, 330, 344, 387, 435, 439–443, 450, 457, 465, 466 v Bybee, Hopper (2001a).

Literatura

Bybee, J., Hopper, P. (2001). Introduction to frequency and the emergence of linguistic structure. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 1–24.

Bybee, J., Hopper, P. (eds.) (2001a). *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 1–24.

Corbett, G., Hippisley, A., Brown, D., Marriott, P. (2001). Frequency, regularity and the paradigm: A perspective from Russian on a complex relation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 201–226.

Fenk-Oczlon, G. (1991). Frequenz und Kognition – Frequenz und Markiertheit. *Folia Linguistica 25*, 361–394.

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 431–448.

Fenk-Oczlon, G. (1990). Frequenz und Kognition – Frequenz und Markiertheit. *Folia Linguistica 25*, 361–394.

Greenberg, J. H. (1966). *Language universals*. The Hague: de Gruyter.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin. Cambridge: The MIT Press.

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology 40*, 1–95.

5.7 FREKVENCE A SLOVOSLED VE FRÁZÍCH

Hypotéza

„Frekventovanější slovo se vyskytuje před méně frekventovaným slovem." (Fenk-Oczlon 2001: 437). Hypotéza se týká frází jako např. v němčině ‚mit Kind und Kegel'.

Postup

Vyhledejte zhruba 500 frází (tj. co možná nejvíce) z frazeologického slovníku a zjistěte, zda je první slovo fráze frekventovanější než druhé. Pro zjištění četnosti slova použijte frekvenční slovník nebo korpus. Proveďte znaménkový test hypotézy.

Kombinujte tuto hypotézu s dalšími, které se týkají frází, a pokuste se najít obecné vysvětlení. Podívejte se na část 2.1, „Behagelův zákon".

Literatura

Chafe, W. (1994). *Discourse, consciousness, and time*. Chicago, London: University of Chicago Press.

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 431–448.

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics 27,* 517–556.

Givón, T. (1984). *Syntax: a functional typological introduction. Volume 1.* Amsterdam, Philadelphia: J. Benjamins.

Givón, T. (1990). *Syntax: a functional typological introduction. Volume 2.* Amsterdam, Philadelphia: J. Benjamins.

Siewierska, A. (1988). *Word order rules*. London, New York, Sydney: Croom Helm.

Sobkowiak, W. (1993). Unmarked-before-marked as a freezing principle. *Language and Speech 36*, 393–414.

5.8 FREKVENCE A KOMPLEXITA FONÉMU

Hypotézy

„…existuje rovnováha mezi velikostí nebo stupněm komplexity fonému a jeho relativní četností výskytu v tom smyslu, že míra nebo stupeň komplexity fonému je inverzní vzhledem k relativní frekvenci jeho výskytu." (Zipf 1935: 49).

„…ve všech případech, kde se dá určit míra stupně komplexity fonémů, je tato míra komplexity v inverzním (ne nutně proporčním) poměru k relativní frekvenci výskytu." (Zipf 1935: 79).

„…zdá se být velmi nepravděpodobné, že by míra komplexity byla příčinou relativní četnosti výskytu. Může být však prokázán opak…" (Zipf 1935: 81).

Postup

Než začnete s testováním dané hypotézy, musíte nejdříve jasně definovat koncept míry komplexity fonému. Následně se na základě spočítaných četností jakéhokoliv fonému pokuste graficky prokázat existenci této závislosti. Ověřte hypotézu na více jazycích, počínaje těmi, které mají v inventáři 13 fonémů, a konče těmi, které jich mají 40.

Podle získaných výsledků buď potvrďte, modifikujte nebo zamítněte danou hypotézu. Porovnejte váš koncept komplexity fonému se Zipfovým pojetím. Specifikujte závislou proměnnou (frekvence nebo komplexita).

Literatura

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin, Cambridge: The MIT Press, 252–258.

5.9 FREKVENCE A FORMA FONÉMU

Hypotézy

„V angličtině existuje jasná korespondence mezi slabými počátečními konsonanty a jejich frekvencí: čím frekventovanější slovo je, tím slabší je jeho počáteční konsonant." (Fenk-Oczlon 2001: 439).

„…v nejvyšší frekvenční třídě se distribuce počátečních konsonantů… výrazně liší od celkové distribuce." „…podíl vlastních konsonantů (obstruentů) je mnohem menší a podíl ostatních fonémů (ne-obstruentů) je mnohem větší než v celé distribuci." (Fenk-Oczlon 2001: 438).

Postup

Fenk-Oczlon považuje méně obstruentní fonémy (např. glidy, vokály) za slabší než ostatní. Definujte jasně rozsah obstruence (míru šumovosti), např. Fenk-Oczlonová škáluje následovně: glidy, likvidy, nazály, frikativy, okluzivy, přičemž vokály jsou nejméně obstruentní, a pokuste se odvodit funkci pro vztah mezi frekvencí každého slova ve frekvenčním slovníku a mírou obstruence jeho prvního fonému. Pokud je hypotéza pravdivá, projeví se velký rozptyl. Zkuste eliminovat tento rozptyl vytvořením frekvenčních tříd. Testujte alespoň prvních 1 000 slov pro zjištění rozdílů mezi frekvencemi slov se slabými a silnými konsonanty. Pokud hypotéza ve vašem jazyce neplatí, zkuste vytvořit jinou hypotézu, tj. zda existuje vztah mezi frekvencí slova a jeho prvním fonémem.

Literatura

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 431–448.

5.10 FREKVENCE A PRODUKČNÍ ÚSILÍ

Hypotéza

„… lze-li říci totéž dvěma způsoby, dostane přednost méně ‚náročný' způsob, což je za normálních okolností kratší a snadněji vyslovitelná varianta." (Dahl 2001: 475).

Postup

Za prvé, pokuste se přesně definovat koncept „jednodušší výslovnosti". Obvykle je spojen s produkčním úsilíma a jeho minimalizací. Za druhé, specifikujte výraz „dostane přednost" v uvedené hypotéze. Snažte se být přesní: znamená to, že se méně náročný způsob vyskytuje častěji než ten „obtížnější"? Nebo že ho nahradí? Za třetí, stanovte, zda hypotéza znamená, že délka a jednoduchost jsou způsobeny frekvencí. Obvykle je frekvence považována za příčinu krátkosti a jednoduchosti (viz část 5.15, „Délka a frekvence"). Uvažujte nad touto hypotézou jako nad problémem nejasné formulace a zkuste ji upřesnit. Rozšiřte váš argument tak, aby se dal testovat. Definujte koncept testovatelnosti a netestovatelnosti hypotézy.

Literatura

Bunge, M. (1967). *Scientific research I*. Berlin, Heidelberg, New York: Springer.

Dahl, Ö. (2001). Inflationary effects in language and elsewhere. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 471–480.

5.11 FREKVENCE A PRODUKTIVITA

Hypotéza

„Čím frekventovanější je nějaký morfém, tím je větší jeho morfologická produktivita." (Krott 2002).

Postup

Jde o zobecnění předchozího problému, jenž se zde vztahuje na jakýkoliv typ morfologického konstruktu (derivace, kompozice, reduplikace). Směr hypotézy, tj. co je závislá a nezávislá proměnná, není pevně stanoven. Excerpujte z korpusu všechny morfémy, zjistěte jejich frekvence a všechny morfologické konstrukce (typy), ve kterých se objevily. Na základě Köhlerova kontrolního cyklu odvoďte teoretickou funkci a zkuste ji aplikovat na empirická data.

Literatura

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Krott, A. (2002). Ein funktionalanalytisches Modell der Wortbildung. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik*. [Dostupné z: http://ubt.opus.hbz-nrw.de/volltexte/2004/279/]

5.12 FREKVENCE A REDUKCE

Hypotézy

„Pravděpodobnostní redukční hypotéza: slovní tvary jsou redukovány, když mají vyšší pravděpodobnost výskytu. Pravděpodobnost slova je podmíněna mnoha aspekty jeho kontextu, včetně sousedních slov, syntaktických

a lexikálních struktur, sémantických očekávání a diskurzních faktorů." (Jurafsky, Bell, Gregory, Raymond 2001: 229).

„…u slov, která jsou silně spjata se sousedními slovy nebo jsou předvídatelná z těchto slov, např. jde o kolokace (sekvence obvykle se spolu-vyskytujících slov), je pravděpodobnější, že u nich dojde k fonologické redukci." (Jurafsky, Bell, Gregory, Raymond 2001: 230).

„…předvídatelnost ovlivňuje nejen délku vokálů, ale má také navíc nezávislý nekategoriální vliv na délku slova." (Jurafsky, Bell, Gregory, Raymond 2001: 239).

„S rostoucí pravděpodobností slova způsobenou sousedním slovem se zkracuje jeho délka." (Jurafsky, Bell, Gregory, Raymond 2001: 240).

„Různá formální slova jakéhokoliv slovníku jsou potom, zdá se, pozůstatky určitých minulých stavů abreviačního procesu. Jsou to jména zkušenostních kategorií, ke kterým je často v řeči odkazováno." (Zipf 1935: 271).

„…když se nějaká smysluplná konfigurace stává relativně frekventovanější, stává se současně méně artikulovanou a více integrovanou." (Zipf 1935: 272).

„…vypouštění […] převažuje spíše u slov s vysokou frekvencí než u slov s nízkou frekvencí . (Pierrehumbert 2001: 138).

„…frekventovanější jednotky mají tendenci redukovat se častěji než jednotky s nízkou frekvencí." (Bush 2001: 257).

„…frekventovaná slova se redukují rychleji než nefrekventovaná slova." (Fenk-Oczlon 2001: 436).

„…zdá se, že procesy zkracování se objevují jako důsledek rostoucí frekvence užití daného slova, ať už v celé komunitě mluvčích, nebo v rámci malých skupin." (Zipf 1935: 33).

„…kdekoliv existuje vztah mezi zkracováním a frekvencí, frekvence je příčinou zkracování." (Zipf 1935: 36).

„…narůstající efekt trvalého zkracování v průběhu evoluce jazyka se částečně odráží ve vztahu mezi frekvencí a délkou dnešních slov." (Zipf 1933: 36).

„…na základě dočasných procesů zkracování není možné statisticky dokázat, že je frekvence nutnou příčinou všech substitucí kratších forem." (Zipf 1933: 37).

Postup

Pokuste se u každého slova oddělit všechny prvky (sousední slova N, syntaktickou strukturu Sy, lexikální strukturu L, sémantické očekávání Se, diskurzní faktory D). Proveďte nezbytná měření a zkuste vyjádřit rozsah redukce jako $R = f(N, Sy, L, Se, D)$. Začněte s lineárním vztahem a postupně jej učiňte komplexnějším, např. $R = f(N)$, $R = f(Sy)$ atd. a kombinujte je. Akceptujte daný vztah jen v případě, že redukuje varianci. Vyjádřete krácení alespoň jako funkci pravděpodobnosti sousedství.

Vytvořte více hypotéz tohoto vztahu a ukažte, že jde o velmi bohatou oblast výzkumu.

Literatura

Bush, N. (2001). Frequency effects and word-boundary palatalization in English. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 255–280.

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 431–448.

Jurafsky, D., Bell, A., Gregory, M., Raymond, W. D. (2001). Probabilistic relations between words: evidence from reduction in lexical production. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 229–254.

Pierrehumbert, J. B. (2001). Exemplar dynamics: word frequency, lenition and contrast. In: Bybee, J., Hopper, P. (eds.), *Frequency and the*

emergence of linguistic structure. Amsterdam, Philadelphia: J. Benjamins, 137–157.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin, Cambridge: The MIT Press, 252–258.

5.13 FREKVENCE A ROZMANITOST

Hypotéza

„…zdá se, že počet různých slov (tj. rozmanitost) je tím větší, čím je nižší frekvence výskytů." (Zipf 1935: 26).

Postup

Hypotéza je velmi jednoduchá. Vyjadřuje, že distribuce frekvencí slov (frekvenční spektrum) monotónně klesá. Hypotéza není specifikována.

Sestavte co nejvíce frekvenčních distribucí slov a pokuste se najít jejich společné rozdělení nebo ukažte, že se ve vašem výběru vyskytuje mnoho různých rozdělení. Nejčastější jde o Zipfovo (zeta) rozdělení, Zipf-Mandelbrotovo rozdělení, Waringovo rozdělení atd. Pokuste se nalézt podmínky, za kterých každé rozdělení platí. Viz také části „Frekvence slov 1, 2, 3" v kapitole 4.

Literatura

Popescu, I.-I., Vidya, M. N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008). *Word frequency studies*. Berlin, New York: de Gruyter.

Laws in Quantitative Linguistics: Word frequency. [online]. Dostupné z: http://lql.uni-trier.de

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin, Cambridge: The MIT Press, 252–258.

5.14 DÉLKA A FREKVENCE

Hypotézy

„…čím je slovo delší, tím méně se používá." (Zipf 1935: 22).

„…velikost slov obecně tíhne k inverznímu (ne nezbytně proporčnímu) vztahu k počtu výskytů." (Zipf 1935: 25).

„…vysoká frekvence je příčinou malé velikosti." (Zipf 1935: 29).

„Bylo zjištěno, že jednoduchý stochastický model umožňuje hrubou predikci výsledků získaných při kombinaci všech slov, nikoliv však v případě, když jsou slova klasifikována jako funkční (synsémantika) a plnovýznamová. Funkční slova jsou krátká a frekvence jejich výskytu je klesající funkcí jejich délky, plnovýznamová slova jsou delší a jejich pravděpodobnost jejich výskytu je relativně nezávislá na délce." [Abstract] (Miller, Newman, Friedman 1958).

„…čím větší je počet tahů [v japonském *kanji*], tím menší je počet výskytů daného slova." (Sanada 2007).

„…délka morfému tíhne k inverznímu poměru k jeho relativní četnosti." (Zipf 1935: 173).

„Míra komplexity morfému je inverzní (ne nezbytně proporční, možná je nějakou nelineární matematickou funkcí) vzhledem k jeho relativní frekvenci." (Zipf 1935: 176).

Výše uvedené názory se liší, tudíž je nezbytné provést důkladný výzkum těchto hypotéz.

Postup

Změřte délku každého slova (slovního tvaru) ve frekvenčním slovníku. Definujte délku jedním ze tří následujících způsobů: (a) počtem fonémů ve slově, (b) počtem slabik ve slově, (c) počtem morfémů ve slově. Určování hranic mezi slabikami nebo morfémy není nutné, stačí určit jejich počet. Následně vezměte slova s frekvencí 1 a vypočítejte jejich průměrnou

délku, pokračujte se slovy s frekvencí 2 a spočítejte jejich průměrnou délku atd. Slova s vysokou frekvencí mohou být sloučena. Pokud je hypotéza pravdivá, bude ve všech případech získána monotónní klesající funkce. Vyjádřete rozdíly mezi jednotlivými křivkami graficky, zkuste odvodit funkce z proporcí. Porovnejte několik jazyků, pokud je to možné, analyzujte silně aglutinační jazyk a zjistěte, zda má silná aglutinace vliv na parametry funkce. Pokud bude mít výsledná funkce oscilační charakter, přečtěte si literaturu uvedenou níže.

Literatura

Baker, S. J. (1951). A linguistic law of constancy: II. *The Journal of General Psychology 44*, 113–120.

Baker, S. J. (1951). Ontogenetic evidence of a correlation between the form and frequency of use of words. *The Journal of General Psychology 44*, 235–251.

Belonogov, G. G. (1962). O nekotorych statističekich zakonomernostjach ruskoj pis´mennoj reči. *Voprosy jazykoznanija 11/1*, 100–101.

Breiter, M. A. (1994). Length of Chinese words in relation to their other systemic features. *Journal of Quantitative Linguistics 1*, 224–231.

Gieseking, K. (2002). Untersuchungen zur Synergetik der englischen Lexik. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik,* 387–433.

Grzybek, P., Altmann, G. (2002). Oscillation in the frequency-length relationship. *Glottometrics 5*, 97–107.

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire. Essai de méthodologie*. Paris: PUF.

Guiter, H. (1977). Les relations /frequence-longeuer-sens/ des mots (langue romanes et anglais). In: *XVI Congresso Internazionale di Linguistica e Filologia Romanza*, Napoli, 15–20 Aprile 1974. Napoli: Macchiaroli, Amsterdam: J. Benjamins, 373–381.

Hammerl, R. (1990). Länge – Frequenz, Länge – Rangnummer. Überprüfung von zwei lexikalischen Modellen. *Glottometrika 12*, 1–24.

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells*. Trier: Wissenschaftlicher Verlag.

Herdan, G. (1966). *The advanced theory of language as choice and chance*. Berlin: Springer.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R., Zörnig, P., Brinkmöller. (1990). Differential equation models for the oscillation of the word length as a function of the frequency. *Glottometrika 12*, 25–40.

Kornai, A. (2002). How many words are there? *Glottometrics 4*, 61–86.

Krott, A. (2002). Ein funktionalanalytisches Modell der Wortbildung. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen in die quantitative und systemtheoretische Linguistik*, 75–126.

Leopold, E. (1997). Frequency spectra within word length classes. In: *Third International Conference on Quantitative Linguistics*, August 26–29, 1997, Helsinki, Finland. Helsinki: Monila, 156.

Leopold, E. (1998). *Stochastische Modellierung lexikalischer Evolutionsprozesse*. Hamburg: Kovač.

Leopold, E. (2000). Length-distribution of words with coinciding frequency. In: *Proceedings of the fourth conference of the International Quantitative Linguistic Association*, Prague, August 24–26. Prague, 76–77.

Miller, G. A., Newman, E. B., Friedman, E. A. (1958). Length-frequency statistics for written English. *Information and Control 1*, 370–389.

Miyajima, T. (1992). Relationship in the length, age and frequency of Classical Japanese words. *Glottometrika 13*, 219–229.

Sanada, H. (1999). Analysis of Japanese vocabulary by the theory of synergetic linguistics. *Journal of Quantitative Linguistics* 6, 239–251.

Sanada, H. (2006). The selection of scholarly terms in basic vocabulary lists. *Goi Kenkyu* (*Studies on vocabulary*) *4*, 21–42.

Sigurd, B., Eeg-Olofsson, M., van de Weijer, J. (2004). Word length, sentence length and frequency Zipf revisited. *Studia Linguistica 58(1)*, 37–52.

Tuldava, J. (1995). *Methods in quantitative linguistics*. Trier: Wissenschaftlicher Verlag.

Universitätsbibliothek Trier. [online]. Dostupné z: http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language.* Cambridge: Harvard University Press.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology.* Boston: Houghton Mifflin, Cambridge: The MIT Press.

5.15 DÉLKA A POLYSÉMIE

Hypotéza

Čím je slovo delší, tím menší je počet jeho významů.

Postup

Tato hypotéza je velmi stará. Obvykle se pracuje s *průměrným* počtem významů pro určitou délku. Výsledkem je obvykle mocninná funkce, která je dobře známá z literatury, dobře ověřena a má velmi obecnou platnost. Nicméně zde stále zůstává nevyřešený problém.

Vytvořte ze slovníku velký výběrový soubor – pokud je to možné, pracujte s celým slovníkem. Změřte délku každého slova ve slabikách (x) a počet jeho významů (y). Jednoznačně definujte, jakým způsobem je měřena druhá proměnná. Potom vytvořte dvojdimenzionální distribuci počtu slov (z) závislých na (x, y), tj. $P(z) = f(x,y)$. Tento problém je obtížný vzhledem k tomu, že není snadné získat výběrový soubor. Pokud nesestavíte vhodný výběrový soubor, použijte data z indonézského slovníku v Altmann et al. (2002: 88). Pokud je to nutné, spojte některé třídy.

Protože prodlužování slova (derivací nebo skládáním) je způsobeno požadavkem specifikace, zvažte možnost považovat počet významů za

nezávislou proměnnou a délku za závislou proměnnou. V tomto případě postačí vyhodnotit jen větší výběrový soubor ze slovníku a aplikovat spojitou funkci. Postupujte podle pravidel synergetické lingvistiky.

Literatura

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie*. Göttingen: Peust & Gutschmidt.

Altmann, G., Beőthy, E., Best, K.-H. (1982). Die Bedeutungsmenge und das Menzerathsche Gesetz. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 35*, 537–543.

Baker, S. J. (1950). The pattern of language. *Journal of General Psychology 42*, 25–66.

Fickermann, I., Markner-Jäger, B., Rothe, U. (1984). Wortlänge und Bedeutungskomplexität. *Glottometrika 6*, 115–126.

Gieseking, K. (2002). Untersuchungen zur Synergetik der englischen Lexik. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik*, 387–433.

Hoffmann, Ch. (2001). Polylexie lexikalischer Einheiten in Texten. In: Uhlířová, L. et al. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs*. Trier: Wissenschaftlicher Verlag, 76–97.

Levickij, V. (2005). Polysemie. In: Altmann, G., Köhler, R., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 458–464.

Sambor, J. (1984). Menzerath´s law and the polysemy of words. *Glottometrika 6*, 94–114.

Universitätsbibliothek Trier. [online]. Dostupné z: http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

5.16 DÉLKA A SLOVNÍ DRUHY 1

Hypotézy

„…příslovce času jsou v průměru méně nezávislá, a proto kratší než příslovce místa." (Zipf 1935: 242).

„Bylo zjištěno, že jednoduchý stochastický model umožňuje hrubou predikci výsledků získaných při kombinaci všech slov, nikoliv však v případě, když jsou slova klasifikována jako funkční (synsémantika) a plnovýznamová. Funkční slova jsou krátká a frekvence jejich výskytu je klesající funkcí jejich délky, plnovýznamová slova jsou delší a jejich pravděpodobnost jejich výskytu je relativně nezávislá na délce." [Abstract] (Miller, Newman, Friedman 1958).

Postup

S pomocí podrobné gramatiky uvažujte všechna časová a prostorová příslovce. Definujte přesně délku. Vypočítejte průměrné délky těchto dvou tříd a porovnejte je pomocí statistického testu. Má Zipf pravdu? Pokud je vytváření výběrového souboru obtížné, použijte jen jednoduchá příslovce, těm složitým se vyhněte.

Literatura

Miller, G. A., Newman, E. B., Friedman, E. A. (1958). Length-frequency statistics for written English. *Information and Control 1*, 370–389.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin, Cambridge: The MIT Press.

5.17 DÉLKA A SLOVNÍ DRUHY 2

Problém

Synsémantická slova se vyskytují častěji než autosémantická slova, tudíž jsou kratší než autosémantika. Proto mají některé slovní druhy kratší průměrnou délku než jiné.

Postup

Seřaďte slova ve frekvenčním slovníku (nebo korpusu) vzestupně podle délky a každé slovo přiřaďte do patřičného slovního druhu. Potom přidělte každému slovu pořadí podle jeho délky. Proveďte (neparametrický) rankový test, abyste dokázali, že jednotlivé slovní druhy mají různé pořadí (tj. že se jednotlivé slovní druhy liší). Jinou možností je vypočítat průměrnou délku všech slov určitého slovního druhu a testovat její rozdíl s jinými slovními druhy.

Literatura

žádná

5.18 DÉLKA VĚTY A DÉLKA KLAUZE

Problém

Testujte Shermanův zákon a Menzerathův zákon na délkách vět a délkách klauzí.

Postup

Spočítejte délky vět v počtu klauzí a délky klauzí v počtu slov v několika textech. Vytvořte jejich frekvenční distribuce (náhodná proměnná je délka).

(1) Prokažte, že se obě distribuce se řídí negativním binomiálním rozdělením. Porovnejte parametry distribucí v různých žánrech a jazycích.

(2) Prokažte, že existuje závislost mezi parametry negativního binomického rozdělení.

(3) Testujte Menzerathovu hypotézu, podle níž platí, že čím delší je věta, tím kratší jsou klauze. Tato závislost má formu mocninné funkce.

(4) Porovnejte různé texty, zkuste najít rozdíly mezi texty a jazyky. Pokud je to možné, zkoumejte především silně aglutinační jazyky. Zkuste najít rozdíly a vysvětlete je.

(5) Zkoumejte vývoj délek vět v určitém typu textů, např. žurnalistické texty v průběhu několika desetiletí.

Literatura

Altmann, G. (1988). Verteilungen der Satzlängen. *Glottometrika 9*, 147–170.

Best, K.-H. (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.

Best, K.-H. (2002). Satzlängen im Deutschen: Verteilungen, Mittelwerte, Sprachwandel. *Göttinger Beiträge zur Sprachwissenschaft 7*, 7–13.

Best, K.-H. (2005). Satzlänge. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 298–304.

Dzhurjuk, T. (2006). Sentence length as a feature of style. *Glottometrics 12*, 55–62.

Heeschen, V. (1994). How long are clauses and sentences in a Papuan language like Eipo? *Semaian 10*, 50–70.

Heups, G. (1983). Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. *Glottometrika 5*, 113–133.

Kaßel, A., Livesey, E. (2001). Untersuchungen zur Satzlängenhäufigkeit im Englischen: Am Beispiel von Texten aus Presse und Literatur (Belletristik). *Glottometrics 1*, 27–50.

Kelih, E., Grzybek, P. (2004). Häufigkeiten von Satzlängen: Zum Faktor der Intervallgröße als Einflussvariable (Am Beispiel slowenischer Texte). *Glottometrics 8*, 23–41.

Niehaus, B. (1997). Untersuchung zur Satzlängenhäufigkeit im Deutschen. *Glottometrika 16*, 213–275.

Teupenhayn, R., Altmann, G. (1984). Clause length and Menzerath´s law. *Glottometrika 6*, 127–138.

Uhlířová, L. (2001). On word length, clause length and sentence length in Bulgarian. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of L. Hřebíček*. Trier: Wissenschaftlicher Verlag, 266–282.

5.19 DÉLKA SLOVA A POLYTEXTUALITA

Hypotéza

Podle hypotézy odvozené od Köhlerova kontrolního cyklu platí, že čím je slovo delší, tím menší je jeho polytextualita, tj. čím je slovo delší, tím menší je počet různých textů, ve kterých se vyskytne.

Postup

Vytvořte několik sad slov různých délek. Do každé sady dejte jen slova stejného slovního druhu. Spočítejte jejich výskyty v jednotlivých textech nějakého korpusu. Zobrazte graficky relaci <*délka, polytextualita*> pro každou sadu. Na základě teoretických předpokladů se následně pokuste odvodit vhodnou funkci. Zjistěte, zda jednotlivé sady (obsahující různé slovní druhy) vykazují odlišné parametry funkce.

Literatura

Gieseking, K. (2002). Untersuchungen zur Synergetik der englischen Lexik. In: Köhler, R. (Hg.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik*, 387–433.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Rothe, U. (1983). Wortlänge und Bedeutungsmenge: Eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen. *Glottometrika 5*, 101–112.

Universitätsbibliothek Trier. [online]. Dostupné z: http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

5.20 DÉLKA SLOV A POZICE VE VĚTĚ

Problém

Někteří badatelé tvrdí, že délka slova se liší v různých pozicích hlavní klauze ve větě. Pokuste se najít obecný vzorec vyjadřující změnu délky slova v jednotlivých pozicích, za hraniční podmínku považujte délku klauze.

Postup

Roztřiďte věty/klauze v dlouhém textu podle jejich délky a vypočítejte pro každou pozici ve větě/klauzi průměrnou délku slova. Vytvořte graf a zkuste najít formální vyjádření vámi vytvořené křivky. Výsledek zobecněte. Nezkoumejte jen indoevropské jazyky, ale vyhněte se monosylabickým jazykům.

Literatura

Behagel, O. (1930). Von deutscher Wortstellung. *Zeitschrift für Deutschkunde 44*, 81–89.

Croft, B. (1981). *Language universals and linguistic typology. Syntax and morphology.* Oxford: Blackwell.

Fenk, A., Fenk-Oczlon, G. (2005). Within-sentence distribution and retention of content words and function words. In: Grzybek, P. (ed.), *Word length studies and related issues*. Dordrecht: Springer, 157–170.

Greenberg, J. H. (1963, 1969). Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J. H. (ed.), *Universals of language. Report of a conference held at Dobbs Ferry,* New York, April 13–15, 1961. 2nd ed. Cambridge: The MIT Press.

Hawkins, J. S. (1983, 1988). *Word order universals*. San Diego: Academic Press.

Hawkins, J. S. (1990). A parsing theory of word order universals. *Linguistic Inquiry 21(2)*, 223–261.

Hawkins, J. S. (1992). Syntactic weight versus information structure in word order variation. In: Jacobs, J. (ed.), *Informationsstruktur und Grammatik*. Opladen: Westdeutscher Verlag, 196–219.

Hawkins, J. S. (1994). *A performance theory of order and constituency.* Cambridge: University Press.

Hoffmann, C. (2002). „Early immediate constitutents" – ein kognitiv--funktionales Prinzip der Wortstellung(svariation). In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen in die quantitative und systemtheoretische Linguistik*, 31–74.

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6*, 46–57.

Niemikorpi, A. (1997). Equilibrium of words in the Finnish frequency dictionary. *Journal of Quantitative Linguistics 4(1–3)*, 190–196.

Siewierska, A. (1993). Syntactic weight vs. information structure and word order variation in Polish. *Journal of Linguistics 29,* 233–265.

Uhlířová, L. (1997). Length vs. order: word length and clause length from the perspective of word order. *Journal of Quantitative Linguistics 4(1–3)*, 266–275.

Uhlířová, L. (1997a). O vztahu mezi délkou slova a jeho polohou ve větě. *Slovo a slovesnost 58*, 174–184.

Uhlířová, L. (1997b). Length vs. order. Word length and clause length from the perspective of word order. *Journal of Quantitative Linguistics 4*, 266–275.

Universitätsbibliothek Trier. [online]. Dostupné z: http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

5.21 DÉLKA SLOVA/MORFÉMU A KOMPOZITA

Hypotéza

„Čím je slovo kratší, tím častěji se objevuje v kompozitech." (Altmann 1989: 104).

Postup

Rozdělte dostatečně velký počet náhodně vybraných substantiv ze slovníku do tříd podle jejich délky (slova ve slabikách, morfémy ve fonémech). Potom pro každé slovo/morfém najděte všechna kompozita, v nichž se vyskytují.

Vypočítejte průměry a vytvořte funkci *průměrná aktivita kompozita = = f(průměrná frekvence)*. Opakujte postup s dalšími slovními druhy.

Literatura

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10*, 100–107.

Krott, A., Schreuder, R., Baayen, R. H. (1999). Complex words in complex words. *Linguistics 37*, 905–926.

Prün, C. (2005). Quantitative Morphologie: Eigenschaften der morphologischen Einheiten und Systeme. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 227–242.

6 Sémantika, synergetika a psycholingvistika

6.1 ABSTRAKTNOST

Problém

Navrhněte způsob měření abstraktnosti textu.

Postup

Nejdříve pracujte jen se substantivy. Pokuste se škálovat jejich abstraktnost (nezaměňovat s obecností) za použití (a) abstraktnosti afixů, (b) typu definice ve výkladovém slovníku, (c) možnosti percepce jejich denotátů. Požádejte probandy, aby provedli hodnocení v určeném intervalu. Nepoužívejte kontextová omezení. Proveďte analogický postup u adjektiv a následně u sloves. Popište přesně vaši škálovací metodu.

Na základě vaší škály abstraktnosti, kterou jste zpracovali pro zkoumané slovní druhy, vytvořte úhrnný index abstraknosti. Zpracujte poetický a vědecký text a vypočítejte míru jejich abstraktnosti.

Pokud jste zdatní ve statistice, zkuste odvodit očekávanou hodnotu a varianci indexu, proveďte asymptotický text významnosti rozdílu dvou textů. Dokažte, že vědecké texty jsou více abstraktní než poetické. Vyhodnoťte mnoho textů a zkuste rozdělit žánry podle míry abstraktnosti.

Literatura

Altarriba, J., Bauer, L. M., Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments & Computers 31*, 578–602.

Gilhooly, K. J., Logie, R. H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behaviour Research Methods and Instrumentation 12*, 395–427.

Groeben, N. (1982). *Leserpsychologie: Textverständniss – Textverständlichkeit*. Münster: Aschendorf.

Paivio, A., Yuille, J. C., Madigan, S. A. (1968). Concreteness, imagery and meaningfulness values of 925 words. *Journal of Experimental Psychology, Monograph Supplement 76*.

Wiemer-Hastings, K., Graesser, A. C. (1998). Abstract noun classification: A neural network approach. *Proceedings of the 20th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, 1036–1042.

Wiemer-Hastings, K., Krug, J., Xu, X. (2006). Imagery, context availability, contextual constraint and abstractness. [Dostupné z: http://conferences.inf.ed.ac.uk/cogsci2001/pdf-files/1106.pdf]

6.2 DISTRIBUCE POLYSÉMIE

Problém

Předpokládá se, že distribuce polysémie má ve slovníku rozdělení, jež se řídí nějakým zákonem. Existují různé modely, v obecné rovině se mluví o Krylovově zákonu. Testujte různé formy tohoto zákona nebo vytvořte vlastní model.

Postup

Na základě studia relevantní literatury uvažujte o problematice týkající se distribuce. Testujte jednotlivé modely na datech vytvořených Steinerovou (1995), která obsahují kompletní Wahrigův německý slovník (který rozlišuje slovní druhy, pokud je to nutné, můžete některé slovní druhy sloučit).

Najděte model, který vykazuje nejlepší shodu s daty. Pokuste se najít argumenty pro jeho opodstatnění.

V článku Levického, Drebeta a Kiika (1999) lze najít distribuce polysémie v němčině (tab. 1, 2, 3). Pokuste se najít společné teoretické rozdělení pro všechna tato data.

Literatura

Krylov, J. K. (1982). Eine Untersuchung statistischer Gesetzmäßigkeiten auf der paradigmatischen Ebene der Lexik natürlicher Sprachen. In: Guiter, H., Arapov, M. V. (eds.), *Studies on Zipf´s law*. Bochum: Brockmeyer, 234–255.

Levickij, V. (2005). Polysemie. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 458–464.

Levickij, V. V., Drebet, V. V., Kiiko, S. V. (1999). Some quantitative characteristics of polysemy of verbs, nouns and adjectives in German. *Journal of Quantitative Linguistics 6(2)*, 172–187.

Steiner, P. (1995). Effects of polylexy on compounding. *Journal of Quantitative Linguistics 2(2)*, 133–140.

6.3 OBEZNÁMENOST A FREKVENCE

Hypotézy

„…výskyty slov jsou vnímány člověkem a […] frekvence slov jsou uchovány v paměti." (Köhler, Rapp 2007).

„…obeznámenost se slovem narůstá relativně s růstem frekvence percepce daného slova…" (Köhler, Rapp 2007).

Postup

Prozkoumejte vztah mezi frekvencí a obeznámeností na datech z jiného jazyka než angličtiny. Od porobandů získejte hodnocení obeznámenosti. Frekvence jednotlivých slov mohou být určeny na základě frekvenčního slovníku nebo korpusu. Potom testujte Köhlerovu-Rappovu hypotézu

$$y = \frac{V}{1 + Ax^b} \ ,$$

kde y je stupeň znalosti, V je maximální hodnota znalosti ve vašich datech (tj. druh empirického limitu), x je frekvence, A a B jsou parametry odhadnuté z vašich dat. B má negativní hodnotu.

Aplikujte funkci na vaše data a vypočítejte determinační koeficient. Porovnejte vaše výsledky s těmi v angličtině. V případě potřeby data vyhlaďte.

Literatura

Kacinik, N., Shears, C., Chiarello, C. (2000). Familiarity for nouns and verbs: not the same as, and better than, frequency. In: Gleitman, L. R., Joshi, A. S. K. (eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Hilsdale, NJ: Erlbaum, 1035.

Köhler, R., Rapp, R. (2007). Familiarity and frequency: a psycholinguistic application of synergetic linguistics. *Glottometrics 15*, 62–70.

Kreuz, R. J. (1987). The subjective familiarity of English homophones. *Memory & Cognition 15*, 154–168.

6.4 OBEZNÁMENOST SE SLANGOVÝMI SLOVY

Problém

Vytvořte škálovací metodu pro měření obeznámenosti se slangovými slovy, jejich sémantické variability a nejednoznačnosti.

Postup

Existují tři možné reakce na otázku „znáte význam tohoto slova?“. Správná odpověď, špatná odpověď a odpověď „nevím“. Vytvořte způsob měření obeznámenosti se slangovými slovy a ptejte se *n* probandů. Vytvořte rankovou frekvenční distribuci významů každého slangového slova a pokuste se nalézt odpovídající rozdělení.

Vypočítejte entropii této distribuce a pokuste se nalézt vztah mezi obeznámeností a entropií polysémie.

Literatura

Altmann, G. (2005). Der Diversifikationsprozess. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 648–659.

Köhler, R., Rapp, R. (2007). Familiarity and frequency: a psycholinguistic application of synergetic linguistics. *Glottometrics 15*, 62–70.

Serdelová, K. (2005). Some properties of slang words. *Glottometrics* 9, 40–45.

6.5 FREKVENCE KANJI

Hypotéza

Čím je slovo frekventovanější, tím dříve je použito ve výuce *kanji* (Sanada 2006), tj. daného čínského znaku.

Postup

Prozkoumejte studijní plány čtení/psaní pro děti na základních školách (v čínštině, japonštině nebo korejštině). Navrhněte způsob měření na časové ose, např. 1, 2, 3,… (první naučené slovo, druhé naučené slovo, třetí naučené naučené slovo), nebo slova naučená za první měsíc, druhý měsíc, třetí měsíc…, nebo slova naučená za první rok, druhý rok… Potom určete

frekvence jednotlivých znaků z frekvenčního slovníku. Pokud máte určeny časové intervaly učení, vypočítejte průměrné frekvence znaků. Zjistěte, zda je pořadí učení nebo čas učení funkcí frekvence. Vytvořte proporční vztah. Kombinujte tento problém s předchozím, zejména uvažujte o *čas učení = f(počet tahů, frekvence)*.

Zkuste najít dvojdimenzionální závislost.

Literatura

Hall, J. E. (1954). Learning as a function of word-frequency. *American Journal of Psychology 67,* 138–140.

Sanada, H. (2006). The selection of scholarly terms in basic vocabulary lists. *Goi Kenkyu* (*Studies on vocabulary*) *4*, 21–42.

6.6 UČENÍ A KOMPLEXITA

Hypotéza

Čím větší je počet tahů v nějakém japonském *kanji*, tím později je toto *kanji* učeno (Sanada 2006).

Postup

Zkoumejte čínské znaky v čínštině, japonštině a korejštině. Děti se ve škole učí každý měsíc určitý počet nových znaků. Vytvořte seznam znaků uspořádaný v pořadí podle toho, kdy se učí, a vypočítejte jejich komplexitu. Aplikujte dva způsoby měření komplexity: (a) počet tahů proti směru psaní, (b) jiné měření komplexity, např. Altmann (2004). Pokuste se vyřešit tři následující problémy.

(1) Zjistěte, zda je možné výše zmíněnou hypotézu vyjádřit odpovídající funkcí v daných jazycích, např. <*počet tahů, pořadí učení*> nebo

<*komplexita, pořadí učení*>. Pokud nelze aplikovat jednoznačnou funkci, vyhlaďte data různými způsoby. Pokuste se nalézt adekvátní funkci.

(2) Pravděpodobně zjistíte, že empirické pořadí nemá příliš plynulý průběh, proto zohledněte školní úrovně. Vypočítejte také variance úrovní a zkuste zjistit, zda existuje relace <*rozptyl komplexity, pořadí učení*>. Tato závislost bude pravdepodobně plynulejší než (1).

(3) Porovnejte tři výše uvedené jazyky a zjistěte, zda jsou závislosti podobné.

Literatura

Sanada, H. (2006). The selection of scholarly terms in basic vocabulary lists. *Goi Kenkyu* (*Studies on vocabulary*) *4*, 21–42.

6.7 UČENÍ SE U DĚTÍ

Problém

Učení se jazyka u dětí je pravidelný proces, který může být zachycen nějakou funkcí. Najděte takovou funkci/funkce.

Postup

Děti se učí různé složky jazyka velmi systematicky. Zkoumejte:

(1) učení se vokálů a konsonantů od prvního do třicátého měsíce života,

(2) učení se nových slov v prvních deseti letech,

(3) prodlužování délky slov (nejen lemmat, ale také slovních tvarů),

(4) vývoj délky vět,

(5) vývoj délky textu,

(6) vývoj rankové frekvenční distribuce slovních tříd.

Pokuste se spojit vaše data do podoby různých sítí a sledujte jejich změny.

Literatura

Jde o obor s velkým množstvím literatury. Nejlépe dostupná je:

Ke, J., Yao, Y. (2008). Analysing language development from a network approach. *Journal of Quantitative Linguistics 15(1)*, 70–99.

6.8 VÝZNAM A FREKVENCE

Hypotéza

„Základní význam nějakého slova je tudíž jeho statisticky nejčastěji se vyskytující význam v dané skupině, pro kterou chceme stanovit tento základní význam." (Zipf 1935: 276).

Postup

Zobecněte tento problém následujícím způsobem. Prokažte, že se jednotlivé významy jakéhokoliv slova řídí rankovým pravděpodobnostním rozdělením, tj. že frekvence jednotlivých významů jsou uspořádány statisticky. Ze slovníku vyberte náhodně nějaká slova s mnoha významy a v nějakém korpusu najděte všechny věty, které daná slova obsahují, abyste mohli dohledat jejich jednotlivé významy ve větách. Jde o jednoduchý problém diverzifikace, někdy označován jako Beöthyové zákon.

Literatura

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*: Berlin, New York: de Gruyter, 646–658.

Baayen, R. H. (2005). Morphological productivity. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*: Berlin, New York: de Gruyter, 243–255.

Paivio, A., Yuille, J. C., Madigan, S. (1968). Concreteness, imagery, and meaningness values for 925 nouns. *Journal of Experimental Psychology Monograph*.

Hay, J. (2003). *Causes and consequences of word structure*. New York: Routledge.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R. (1991). Diversification of coding methods in grammar. In: Rothe, U. (ed.), *Diversification processes in language: grammar.* Hagen: Rottmann, 47–55.

Reder, L. M., Anderson, J. R., Bjork, R. A. (1974). A semantic interpretation ov encoding specificity. *Journal of Experimental Psychology 102*, 648–656.

Rothe, U. (ed.) (1991). *Diversification processes in language: grammar.* Hagen: Rottmann.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology.* Boston: Houghton Mifflin, Cambridge: The MIT Press, 252–258.

6.9 INVENTÁŘ MORFÉMŮ A JEJICH POLYSÉMIE

Hypotéza

Čím větší je průměrná polysémie morfémů, tím menší je inventář morfémů v jazyce (Krott 2002: 77f).

Postup

Tato hypotéza vyplývá ze Zipfovy hypotézy o rovnováze mezi plnou a nulovou polysémií. Oba extrémy jsou nemožné. Není snadné tuto hypotézu

testovat: je potřeba analyzovat alespoň 10 jazyků, aby bylo možné sledovat průběh funkce. Rozptyl bude pravděpodobně ohromný. Bylo by vhodné začít od izolovaných jazyků směrem k jazykům moderních civilizací. Je nezbytné mít k dispozici kompletní seznam morfémů/morfů a také tým specialistů.

Literatura

Krott, A. (2002). Ein funktionalanalytisches Modell der Wortbildung. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen zur quatitativen und systemtheoretischen Linguistik*. http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

Prün, C., Steiner, P. (2005). Quantitative Morphologie: Eigenschaften der morphologischen Einheiten und Systeme. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 227–242.

6.10 MORFOLOGIE VS. FONOLOGIE

Hypotéza

Vzhledem k tomu, že jsou koncovky obvyklé krátké (často se vytvořily z nezávislých slov), některé jejich fonémy byly odstraněny, obvykle vokály. Proto se můžeme ptát: existuje vzájemný vztah mezi formováním (souhláskových) shluků a rozsahem flexe (aglutinace) v jazyce? (Skalička 1964).

Postup

Vytvořte tabulku jednotlivých souhláskových shluků v textech alespoň u 10 různých jazyků z rozdílných jazykových rodin. Aplikujte asociační měření (např. Harary, Paper 1957) k vyjádření „tendence ke shlukování". Změřte rozsah flexe nebo aglutinace (formování afixů). Aplikujte Greenbergovy/Krupovy indexy a porovnejte výsledky v různých jazycích.

(1) Najděte vztah mezi shlukováním a flexí/aglutinací.

(2) Definujte nový index flexe/aglutinace.

(3) Pokuste se najít vztah mezi vytvářením shluků s ostatními vlastnostmi jazyka.

Literatura

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics 26*, 178–194.

Harary, F., Paper, H. H. (1957). Toward a general calculus of phonemic distribution. *Language 33*, 143–169.

Krupa, V. (1965). On quantification of typology. *Linguistics 12*, 31–36.

Skalička, V. (1964). Konsonantenkombination und linguistische Typologie. *Travaux linguistiques de Prague 1*, 111–114.

6.11 INVENTÁŘ FONÉMŮ VS. DÉLKA MORFÉMŮ

Hypotéza

Čím větší je inventář fonémů v jazyce, tím kratší jsou jeho morfémy (viz Hockett 1968: 93).

Postup

Zvolte si tři jazyky s velmi rozdílnou velikostí inventáře fonémů a použijte již připravené zdroje dat (např. Karpilovska [2002] pro ukrajinštinu, kde najdete všechny kořeny slov tohoto jazyka). Vypočítejte průměrnou délku morfému u vybraných jazyků, použijte náhodné výběrové soubory o velikosti kolem 500 morfémů. Zjistěte, zda jsou tyto délky stejné, nebo se mění v závislosti s narůstajícím inventářem fonémů. Vytvořte hypotézu a ověřte ji na více jazycích.

Literatura

Hockett, Ch. F. (19683). *A course in modern linguistics*. New York: McMillan.

Karpilovska, E. A. (2002). *Korenevij gnizdovij slovnik ukrainskoi movi*. Kiiv: Ukrains´ka enciklopedija.

6.12 POLYSÉMIE A SKLÁDÁNÍ SLOV

Hypotéza

„Čím větší je polylexie slova, tím více kompozit dané slovo tvoří." (Rothe 1988).

Postup

Testujte hypotézu za použití výkladového slovníku a systematicky vytvořeného výběrového souboru 1 000 slov. Definujte přesně polysémii a pojem kompozita ve vámi zkoumaném jazyce. Pro každé slovo určete počet jeho významů (podle slovníku) a počet kompozit, které tvoří. Prokažte, že závislost $počet\ kompozit = f(počet\ významů)$ monotónně roste. Najděte vhodnou funkci a testujte její adekvátnost.

Literatura

Hammerl, R. (1990). Überprüfung einer Hypothese zur Kompositabildung (am polnischen Sprachmaterial). *Glottometrika 12*, 73–83.

Rothe, U. (1988). Polylexy and compounding. *Glottometrika 9,* 121–134.

6.13 SÉMANTICKÉ TŘÍDY

Problém

Řídí se sémantická klasifikace slov rankovým frekvenčním rozdělením?

Postup

Hypotéza v kvantitativní lingvistice říká: pokud je nějaká lingvistická kategorie konstruována „přirozeně", potom se distribuce jejích prvků řídí rankovým frekvenčním rozdělením Zipfova typu. Testujte tuto hypotézu na datech Levického a Lučaka (2005, tab. 7, s. 223, poslední dva sloupce). Tito autoři navrhli 20 typů sloves a prezentovali jejich frekvence pro angličtinu.

Následně změňte pořadí ve sloupcích v tabulce podle pořadí v předposledním řádku tabulky. Pokuste se najít dvojdimenzionální rankovou frekvenční distribuci pro tuto klasifikaci tříd a času. Pokud nebudete úspěšní, najděte alespoň korelaci mezi těmito dvěma klasifikacemi.

Literatura

Levickij, V., Lučak, M. (2005). Category of tense and verb semantics in the English language. *Journal of Quantitative Linguistics 12(2–3)*, 212–238.

6.14 SÉMANTICKÁ DIVERZIFIKACE

Hypotéza

Jednotlivé významy jakéhokoliv slova se vyskytují s různou frekvencí. Tyto frekvence seřazené sestupně se řídí přímým patřičným rankovým frekvenčním rozdělením.

Postup

Tento fakt je známý z mnoha publikací. Zde je vaším úkolem sledovat distribuce jednotlivých slovních druhů. Vytvořte výběrový soubor slov sestávající, řekněme, z 5 substantiv, 5 sloves, 5 adjektiv, 5 předložek atd., a spočítejte frekvence jednotlivých významů všech zkoumaných slov. Vytvořte empirickou rankovou frekvenční distribuci a pokuste se pro ni najít

vhodné modely. Kromě toho dokažte, že předložky jsou více diverzifikovány než substantiva atd. Prokažte, že jazyky se sice zásadně liší v sémantické diverzifikaci, ale řídí se stejnými modely. Na základě teoretických argumentů navrhněte nové modely.

Analyzujte kompletní inventář spojek ve vašem jazyce a vyhodnoťte výsledky.

Literatura

Altmann, G. (1985). Semantische Diversifikation. *Folia Linguistica 19,* 177–200.

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 646–657.

Altmann, G., Best, K.-H., Kind, B. (1987). Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. *Glottometrika 8,* 130–139.

Beőthy, E., Altmann, G. (1984). Semantic diversification of Hungarian verbal prefixes. III. „föl-", „el-", „be-". *Glottometrika 7,* 45–56.

Rothe, U. (1986). Die Semantik des textuellen et. Frankfurt am Main: Lang.

Rothe, U. (ed.) (1991). *Diversification processes in language: grammar.* Hagen: Rottmann.

7 Typologie

7.1 ENTROPIE A SYNTETISMUS

Problém

Má entropie frekvencí slov něco společného se syntetičností jazyka?

Postup

Vypočítejte frekvenční spektra slovních tvarů v několika textech (frekvence f_x je počet slov v textu, která se v něm vyskytla právě x krát). Potom vypočítejte entropii podle rovnice v části 8.10, „Index opakování a entropie“. Spočítejte průměrné entropie všech textů a porovnejte výsledky s následující tabulkou.

TABULKA 7.1.1
Průměrné entropie frekvenčního spektra slov ve 20 jazycích

jazyk	průměr
maďarština	0,9577
latina	1,2203
němčina	1,2980
rumunština	1,3252
bulharština	1,3279
čeština	1,3510
ruština	1,5145

jazyk	průměr
italština	1,5325
tagalog	1,5721
indonézština	1,5823
slovenština	1,6344
maráthština	1,6532
lakotština	1,8002
kannadština	1,8683
angličtina	2,2791
rarotongština	2,6337
samojština	2,7099
maorština	2,7696
markézština	2,8490
havajština	2,8946

Z tabulky je patrné, že čím analytičtější je jazyk, tím je větší entropie frekvenčního spektra slov. Do kterého intervalu patří vaše výsledky? Pokuste se vysvětlit tento jev. Vypočítejte také indexy opakování slov ve vašich textech a porovnejte je s uvedenou literaturou. Srovnejte své výsledky s hodnotami v části 7.8, „Syntetismus v jazyce“.

Literatura

Popescu, I.-I., Altmann, G. (2007). On diversity of word frequencies and language typology. *Göttinger Beiträge zur Sprachwissenschaft 14*, 81–91.

7.2 HOMONYMIE A SYNONYMIE AFIXŮ 1

Problém

Pokud je v nějakém jazyce deklinační systém (např. latina, slovanské jazyky), pak jsou některé deklinační afixy homonymní (stejná forma, ale jiný význam/funkce) a některé synonymní (jiná forma, ale stejný význam/kategorie).

Postup

(1) Pokuste se kvantitativně vyjádřit rozsah homonymie a synonymie v nějakém jazyce.

(2) Prokažte, že homonyma nejsou zastoupena se stejnou frekvencí (např. testujte homogenitu).

(3) Prokažte, že synonyma nejsou v rámci dané kategorie zastoupena stejnou frekvencí (např. testujte homogenitu).

(4) Vytvořte empirickou frekvenční distribuci některých kategorií a prozkoumejte vztah mezi frekvencí pádu a průměrnou délkou afixů.

(5) Sledujte další postup v následující části 7.3, „Homonymie a synonymie afixů 2“.

Literatura

Skalička, V. (2005–2006). *Souborné dílo I–III.* Praha: Nakladatelství Karolinum.

7.3 HOMONYMIE A SYNONYMIE AFIXŮ 2

Hypotéza

Synonymní afixy se řídí svým vlastním rankovým frekvenčním rozdělením.

Postup

Vytvořte seznam všech afixů v nějakém jazyce (jak derivačních, tak flektivních) a jim odpovídajících kategorií, např. v angličtině *-s* vyjadřuje genitiv u substantiv, plurál substantiv, 3. os. sg. sloves; anglické *-ity* a další afixy vyjadřují abstraknost atd. Ignorujte fakt, že genitiv může mít velké množství různých významů, zaměřte se jen na kategorie. Seznam by měl mít podobu tabulky s afixy v prvním sloupci a kategoriemi (významy) v dalších sloupcích. Označte ty buňky tabulky (řádek), u nichž má afix odpovídající význam (sloupec). Do těchto buněk zaneste frekvence jednotlivých afixů zjištěných na základě korpusu nebo frekvenčního slovníku.

Seřaďte řádky podle frekvencí afixů. Potom spočítejte počet označených buňek v každé kategorii (sloupci)a spočítejte počet afixů, které se pojí s danou kategorií. Seřaďte sloupce podle těchto hodnot.

Pro každý jednotlivý řádek a sloupec nalezněte odpovídající rankové frekvenční rozdělení. Použijte taková rozdělení, která nemají více než dva parametry, protože řádky a sloupce jsou krátké. Zkoumejte chování parametrů. Nakonec se pokuste nalézt dvojdimenzionální rozdělení pro celou tabulku. Interpretujte výsledky. Považujte toto rankové frekvenční rozdělení za jediné kritérium „správnosti" ve vašem seznamu afixů a jim přiřazených kategorií. Pokud nebudete s výsledky spokojeni, zaměřte se na jiné gramatické popisy daného jazyka. Na druhou stranu, odchylka může být projevem začínající samoorganizace (opouštění rovnováhy) nebo projevem vlivu samoregulace (obnovení rovnováhy). V každém jazyce budou nějaké „výjimky", jež jsou projevem dynamiky jazyka.

Literatura
žádná

7.4 FLEXE OBECNĚ

Problém
Navrhněte různé způsoby měření míry flexe v jazyce. Vytvořte odlišné způsoby měření pro gramatiku (*langue*) a pro text (*parole*). Pokuste se vypočítat rozdíl v míře flexe mezi psanou a mluvenou francouzštinou na základě analýzy textů.

Porovnejte jednotlivé případy a spočítejte vývoj ztráty flexe v mluvené francouzštině. Porovnejte starou angličtinu s moderní angličtinou, latinu se španělštinou, starou ruštinu s moderní ruštinou. Nejprve aplikujte indexy, které navrhl Greenberg a Krupa, ale pokuste se také definovat nějaké nové indexy.

Hypotéza
„…čím větší je počet různých flektivních afixů v jazyce, tím úměrně menší bude počet různých kořenů, které se vyskytnou v proudu řeči, ve srovnání s počtem různých slov vytvořených z těchto kořenů." (Zipf 1935: 252–253).

Postup
(1) Přesně definujte pojem flexe. Pokud jde o Greenbergův index, vytvořte velký výběrový soubor z korpusu, spočítejte počet všech slov a počet slov s flexí. Jejich poměr vyjadřuje míru flexe. Tento poměr považujte za proporci, kterou je možné zpracovat statisticky. Porovnejte několik jazyků.

(2) Pokuste se nalézt funkci, která by reprezentovala Zipfovu hypotézu. Podle Zipfovy teorie distribuce slov by mělo platit $y = k/x^2$. Ověřte vhodnost této funkce, a pokud to bude nezbytné, vytvořte novou teorii.

Literatura

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics* 26, 178–194.

Krupa, V. (1965). On quantification of typology. *Linguistics 12*, 31–36.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin, Cambridge: The MIT Press, 252–258.

7.5 DÉLKA MORFU

Hypotéza

Podle Skaličky (2006: 988, 1054) je délka morfu jedním z ukazatelů jazykové typologie. To znamená, že „v jazycích s vysokým stupněm polysyntetičnosti jsou morfy krátké."

Postup

Ověřte tento předpoklad na textech několika jazyků. Nejprve definujte způsob měření délky morfu (v počtu fonémů nebo slabik), rozhodněte, zda by měl být započítáván nulový morfém. Následně přepište text na morfémy a vytvořte distribuci jejich délek. Best (2005) uvádí, že by měla být data distribuována podle hyper-Poissonova rozdělení. Testujte tuto hypotézu a sledujte rozdíly mezi parametry v jednotlivých jazycích. Zkoumejte několik textů ve zkoumaných jazycích a testujte homogenitu.

Nicméně třída s nejmenším počtem prvků může být projevem charakteristické chování v určitých jazycích. Považujte relativní frekvenci nejmenší třídy ($[x = 0]$ nebo 1, v závislosti na způsobu měření) za charakteristiku jazyka a pokuste se najít další měřitelnou vlastnost, která s ní souvisí.

Literatura

Best, K.-H. (2005). Morphlänge. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handboook*. Berlin, New York: de Gruyter, 255–260.

Skalička, V. (2006). *Souborné dílo III*. Praha: Nakladatelství Karolinum. [zejména články: „Zum Problem des Donausprachbundes". *Ural-altaische Jahrbücher 40 (1–2)*, 1968, 1–9 a „K voprosu o tipologii". *Voprosy jazykoznanija,* 1966, 22–30.]

7.6 POPESCŮV TYPOLOGICKÝ INDIKÁTOR A

Problém

Popescův indikátor *a* vyjadřuje stupeň syntetičnosti jazyka. Porovnejte jazyk, který zkoumáte, s níže uvedenou tabulkou a s obvyklými indexy syntetismu.

Postup

Spočítejte frekvence různých slov (ne lemmat, ale slovních tvarů) v textu. Určete *h*-bod (viz kap. 4 „Textologie") následujícím způsobem: (a) *h*-bod je takový bod, v němž *rank = frekvence*. (b) Pokud takový bod nelze určit, aplikujte vzorec

$$C = \frac{1}{f_r - r} \ ,$$

kde f_r je frekvence v pořadí r a r je dané pořadí slova. C roste až do bodu zlomu, kde začíná být negativní a opět roste. Spojte největší pozitivní C s nejmenším negativním C přímkou. Průsečík této přímky s osou x je h-bod (Popescu et al. 2008).

Uvažujte délku textu N a pro výpočet a použijte vzorec

$$a = \frac{N}{h^2} \ .$$

Analyzujte co nejvíce textů daného jazyka a vypočítejte průměrné a. Získanou hodnotu zaneste do tabulky (viz níže).

Porovnejte vaše průměrné a s ostatními jazyky v tabulce za použití testu

$$t = \frac{|\overline{a_1} - \overline{a_2}|}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \ ,$$

kde

$$s^2 = \frac{\sum_{i=1}^{n_1}(a_{i1} - \overline{a_1})^2 + \sum_{i=1}^{n_1}(a_{i2} - \overline{a_2})^2}{n_1 + n_2 - 2}$$

a t má $n_1 + n_2$ - 2 stupňů volnosti. Hodnoty jednotlivých a najdete u Popesca et al. (2008, tabulka 3.1.1).

Další testy pro měření rozdílů mezi dvojicemi textů najdete u Popesca et al. (2008).

TABULKA 7.6.1

Průměrné hodnoty a ve 20 jazycích (Popescu et al. 2008)

jazyk	$\overline{a}$	*n*	jazyk	$\overline{a}$	*n*
samoánština	4.56	5	italština	8.41	5
rarotonganština	5.02	5	rumunština	9.15	6
havajština	5.37	6	slovinština	9.19	5

jazyk	$\overline{a}$	*n*	jazyk	$\overline{a}$	*n*
maorština	5.53	5	indonézština	9.58	5
lakotština	5.69	4	ruština	10.10	5
markézština	5.69	3	čeština	10.33	10
tagalog	7.24	3	maráthština	11.82	50
angličtina	7.65	13	kannadština	16.58	47
bulharština	7.81	10	maďarština	18.02	5
němčina	8.39	17	latina	19.56	6

Literatura

Popescu, I.-I., Vidya, M. N., Uhlířová, L., Pustet, R., A., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008). *Word frequency studies*. Berlin, New York: de Gruyter.

7.7 DÉLKA KOŘENU A ROZSAH DERIVACE

Hypotéza

Když se podíváme na výše zmíněný Skaličkův systém, můžeme předpokládat, že pokud má jazyk krátké kořeny (v průměru), má bohatou derivaci.

Postup

Nejdříve definujte způsob měření délky kořenu (např. v počtu fonémů), následně definujte rozsah derivace. Můžete použít Greenbergův-Krupův index. Tento problém můžete řešit na základě slovníku nebo korpusu. Abyste získali empirickou formu závislosti, musí být analyzováno několik jazyků, popř. můžete použít výsledky z typologické literatury. Na vyšší úrovni výzkumu by měla být závislost odvozena teoreticky. Pokud neplatí,

omezte ceteris paribus podmínku a hledejte třetí proměnnou, nebo se pokuste najít hraniční podmínky.

Literatura

Skalička, V. (2005–2006). *Souborné dílo I–III*. Praha: Nakladatelství Karolinum.

7.8 SYNTETISMUS V JAZYCE

Problém

Jazyky s mnoha afixy a flexí jsou silně syntetické. Pokuste se navrhnout způsob měření syntetismu a aplikujte jej na několik jazyků.

Postup

Začněte s definovaním míry syntetičnosti prostřednictvím nekořenových morfémů, které naleznete v korpusu (jen typy). Potom se ji pokuste definovat prostřednictvím distribuce slov s 0, 1, 2,… afixy, mezi něž se řadí prefixy, infixy, sufixy a cirkumfixy. Pokuste se navrhnout jiné způsoby měření syntetismu. Nakonec ověřte, zda vysoký stupeň syntetičnosti souvisí s průměrnou délkou slov.

Začnete s čtyřmi definicemi:

(1) W/M (W = *počet slov*, M = *počet morfémů*).

(2) R/M (R = *počet kořenových morfémů*).

(3) S/W (S = *počet vět*).

(4) L/V (L = *počet lexémů/lemmat*, V = *počet slovních tvarů*).

Zjistěte, zda jsou výsledky těchto měření stejné. Analyzujte několik krátkých textů v každém jazyce.

Pokud je to možné, odvoďte nějaké statistické vlastnosti těchto nebo nových indexů. Najděte jiné vlastnosti, které jsou spojeny se syntetismem. Testujte, jestli $L = aV^b$ (Tuldava 1995: 154).

Literatura

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics* 26, 178–194.

Kelemen, J. (1970). Sprachtypologie und Sprachstatistik. In: Dezsö, L., Hajdú, P. (eds.), *Theoretical problems of typology and the Northern Eurasian languages*. Amsterdam, 53–63.

Krupa, V. (1965). On quantification of typology. *Linguistics 12*, 31–36.

Slavíčková, E. (1968). Toward a typological evaluation of related languages. *Travaux linguistiques de Prague 3*, 281–298.

Tuldava, J. (1995). The ratio of word forms and lexemes in texts. In: Tuldava, J. (1995), *Methods in quantitative linguistics*. Trier: Wissenschaftlicher Verlag, 151–159.

7.9 VOKALICKÝ JAZYK

Problém

Existují různé názory na počet vokálů a konsonantů v jazyce. Pokuste se navrhnout jasnou definici vokaličnosti.

Postup

Uvažujte různé způsoby měření „vokaličnosti“. Některé z nich testujte v různých jazycích za použití jak inventářů, tak korpusů. Pokuste se najít vztah mezi některým z těchto měření a nějakou jinou vlastností jazyka, např. stupněm flexe.

Literatura

Altmann, G., Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. München: Fink.

7.10 DÉLKA SLOVA A KONGRUENCE

Hypotéza

S růstem míry kongruence v textu narůstá průměrná délka slova (Skalička 2005–2006).

Postup

Toto je jedna z možných hypotéz odvozených ze Skaličkova systému. Kongruence je obvykle reprezentována určitými (derivačními nebo flektivními) afixy. Tyto afixy prodlužují slovo, zvláště v jazycích se silnou aglutinací.

(1) Analyzujte 10–20 textů z jednoho jazyka. Vypočítejte průměrnou délku slova a proporce slov, u nichž se projevuje kongruence. Počítejte jen explicitní kongruenci, např. v němčině *zu diesen schönen Häusern* jsou spojena tři slova shodou, ale v angličtině *to these nice houses* pouze dvě slova, v indonézštině (*kepada rumah-rumah bagus ini*) není žádné. Slovo může mít zároveň více případů kongruence. Definujte přesně přítomnost takového případu. Potom se pokuste prokázat, zda existujte relace <*shoda, délka slova*>.

(2) Proveďte stejný postup na datech z 10 různých jazyků (nejen indoevropských) a zjistěte, jestli uvedený vztah platí. Pokud je hypotéza pravdivá, pokuste se vytvořit funkci vyjadřující tento vztah. Pokud se takový vztah neprojeví, pokuste se najít další proměnné, které vedou k tomuto vztahu.

Literatura

Skalička, V. (1966). Ein „typologisches Konstrukt". *Travaux linguistiques de Prague 2*, 157–163.

Skalička, V. (2005–2006). *Souborné dílo I–III.* Praha: Nakladatelství Karolinum.

7.11 POŘADÍ SLOV A FLEXE

Hypotéza

„Čím větší je míra flexe v jazyce, tím volnější je jeho slovosled. Přítomnost nebo nepřítomnost prvků flexe a míra jejich použití modifikují povahu syntaktického uspořádání." (Zipf 1935: 246).

Postup

Navrhněte metodu pro měření míry volnosti slovosledu ve větě. Použijte Greenbergovy nebo Krupovy indexy (nebo navrhněte vlastní) pro měření míry flexe. Pokuste se vyjádřit relaci *volnost slovosledu = f(míra inflexe)* ve formě funkce. Testujte adekvátnost této funkce.

Literatura

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics 26*, 178–194.

Krupa, V. (1965). On quantification of typology. *Linguistics 12*, 31–36.

Skalička, V. (1966). Ein „typologisches Konstrukt". *Travaux linguistiques de Prague 2*, 157–163.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology.* Boston: Houghton Mifflin, Cambridge: The MIT Press.

8 Obecné problémy

8.1 DISTRIBUCE

Problém

Najděte empiricky vhodné rozdělení pro různé jednotky v textu.

Postup

R = ranková frekvenční distribuce, F = frekvenční spektrum nebo *frekvence četností.*

Zabývejte se následujícími jednotkami: fonémy (*R*), písmena (*R*), grafémy (*R*), slabiky (*R*), délky slabiky (*F*), slova (*F*), délka slova (*F*), délka klauze (*F*), délka věty (*F*), slovní třída (*R*), počet významů slova (podle slovníku) (*F*). Najděte „nejlepší" rozdělení, sledujte jejich formy a parametry. Pokud je to možné, pokuste se nalézt společný základ pro všechna *R* a *F*. Charakterizujte text jako vektor vlastností distribuce.

Literatura

Grzybek, P. (2006). History and methodology of word length studies. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues*. Dordrecht: Springer, 15–90.

Wimmer, G., Altmann, G. (2006). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues*. Dordrecht: Springer, 329–337.

8.2 ENTROPIE A VELIKOST INVENTÁŘE

Problém

Vzhledem k tomu, že jazykové jednotky mají velmi charakteristické rankové frekvenční distribuce nebo frekvenční spektra, musí platit, že entropie těchto distribucí je přímo závislá na velikosti inventáře. Ve fonémice (distribuce fonémů nebo písmen) byla již tato závislost prokázána.

Postup

Pokuste se prokázat, že má tato hypotéza širší rozsah. Připravte data s frekvencemi fonémů/písmen z jazyků s různou velikostí inventáře a z různě dlouhých textů. Vypočítejte frekvenční distribuce slov, velikost inventářů (inventář typů nebo tokenů) a entropii. Zkoumejte závislost entropie na velikosti inventáře. V prvním kroku omezte svou analýzu jen na empirická šetření. V druhém kroku se pokuste nalézt pro vaše data vhodné teoretické rozdělení a odvoďte entropii teoreticky. Následne se pokuste prokázat, že se empirické entropie řídí teoretickou funkcí.

Literatura

Altmann, G., Lehfeldt, W. (1980). *Einführung in die quantitative Phonologie*. Bochum: Brockmeyer.

8.3 APLIKACE TEORETICKÉHO ROZDĚLENÍ

Problém

Explorativní aplikace teoretického rozdělení na data je jen prvním krokem, nikoliv posledním.

Postup

V příloze článku Galea a Sampsona (1995: 237) je kompletní frekvenční spektrum kanonických tvarů.

(1) Použijte nějaký počítačový program k nalezení všech rozdělení, která empiricky odpovídají těmto datům.

(2) Vytvořte empirickou rankovou frekvenční distribuci těchto dat a najděte rankové frekvenční rozdělení.

(3) Vysvětlete, proč je koncept „neviditelných druhů" v lingvistice špatný.

Pokud pro data z výše uvedeného článku naleznete vhodný vzorec, pokuste se jej odvodit z teoretických předpokladů. Prokažte, že pro daná data vždy existuje několik „dobrých" rozdělení. Přednost musí dostat vždy takové rozdělení, které je odvozeno z teoretických předpokladů.

Referenc

Gale, W. A., Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics 2(3)*, 217–227.

8.4 VYTVÁŘENÍ HYPOTÉZ POMOCÍ FAKTOROVÉ ANALÝZY

Problém

Faktorová analýza může pomoci k získání množiny vlastností, které jsou vzájemně nějak propojeny. Tato propojení v rámci daného faktoru jsou zdrojem pro tvoření hypotéz. Pokuste se nalézt a odvodit hypotézy.

Postup

Prostudujte nejdříve článek Tuldavy (1995a: 84), kde je možné najít některé společné vlastnosti. Vzhledem k tomu, že v článku nejsou uvedena konkrétní data, proveďte ve vašem jazyce textovou analýzu 12 vlastností podle Tuldavy. Potom vytvořte hypotézy týkající se vzájemných závislostí mezi těmito vlastnostmi. Pokuste se pro každý faktor navrhnout köhlerovský samoregulační cyklus. Definujte každou kategorii v logaritmicko-lineární formě, vytvořte vzorce. Nejdříve kombinujte dvojice vlastností, potom trojice, čtveřice atd., tj. pojímejte faktor jako multidimenzionální strukturu. Nakonec porovnejte výsledky z několika jazyků, tj. zkontrolujte, zda vámi vytvořený kontrolní cyklus platí. Tuldava použil následující vlastnosti: počet substantiv, adjektiv, zájmen, sloves, adverbií, pre/postpozicí, spojek, autosémantik, koncentraci funkčních slov, entropii, frekventované tvary slov a ojedinělá slova. Uvažujte nad dalšími vlastnostmi.

Literatura

Tuldava, J. (1995a). An attempt at quantitative analysis of the style of fiction. In: Tuldava, J., *Methods in quantitative linguistics*. Trier: Wissenschaftlicher Verlag, 73–92.

Tuldava, J. (1995b). A comparison of subjective and objective characteristics ov style. In: Tuldava, J., *Methods in quantitative linguistics*. Trier: Wissenschaftlicher Verlag, 93–108.

8.5 IKONIČNOST

Problém

Existuje obrovské množství prací o ikonech, indexech a symbolech. Ve všech jazycích je možné najít enormní množství všech těchto znaků. Ikoničnost v jednotlivých jazycích je popsána v několika knihách. Bohužel neexistuje metoda pro měření rozsahu ikoničnosti, indexovosti

a symboličnosti jednotlivých znaků. Proto by bylo pro sémiotiku velmi přínosné vytvořit kvantifikaci těchto vlastností, což by usnadnilo hledání vztahů mezi těmito a dalšími vlastnostmi jazyka.

V jazyce existují slova ikonického původu, která dnes mají status symbolu. Přeměna ikonu v symbol však není náhlá. Postupně dochází ke ztrátě inkonicity a nárustu symboličnosti. Pokuste se navrhnout nějakou metodu modelující tento proces.

Literatura
žádná

8.6 TVOŘENÍ INDEXŮ

Problém
Tvoření indexů je složitý proces. Nestačí pouze navrhnout poměry některých kvantit. Je nutné uvést všechny vlastnosti indexu.

Postup
Mikk (1997) vytvořil *index komplikovanosti slovních druhů* pro analýzu porozumění textu: $WCC = (N + Adj)/(V + Adj)$. Pokuste se interpretovat tento index a zjistěte obor hodnot *WCC*. Pokud ho nenajdete, změňte index tak, aby výsledky ležely v intervalu <0, 1> a tento nový index interpretujte. Najděte očekávanou hodnotu a variance nového indexu. Vytvořte asymptotický test, který umožní porovnávat dvojice textů.

Tuldava a Vilup (1976: 94) navrhli index substantivity = N/V. Proveďte stejnou analýzu jako u *WCC*.

Viz také část 4.24, „Poměry".

Literatura

Altmann, G. (1978). Zur Anwendung der Quotiente in der Textanalyse. *Glottometrika 1*, 91–106.

Mikk, J. (1997). Parts of speech in predicting reading comprehension. *Journal of Quantitative Linguistics 4(1–3)*, 156–163.

Tuldava, J., Villup, A. (1976). Sõnaliikide sagedusest ilukirjandusproosa autorikõnes. *Töid keelestatistika alalt 1*, 61–102. [with English summary].

8.7 MENZERATHŮV ZÁKON

Problém

Podle Menzerathova zákona platí následující tvrzení: čím delší je konstrukt, tím kratší jsou jeho konstituenty.

Postup

Ověřte hypotézu v jazyce, který nebyl dosud takto analyzován. Počítejte (a) délku věty (v počtu klauzí) vs. délku klauze (v počtu slov), (b) délku slova (v počtu slabik) vs. délku slabiky (v počtu fonémů), (c) délku slova (v počtu slabik) vs. délku morfu (v počtu fonémů), (d) délku slova (v počtu slabik) vs. trvání slabiky (v milisekundách).

Pokuste se analyzovat silně aglutinační jazyk. Pokud výsledky nepotvrdí hypotézu, vysvětlete proč.

Zkoumejte vztah délky věty vs. délky slova, jež je známý jako Arensův zákon. Výsledky interpretujte.

Literatura

Cramer, I. M. (2005). Das Menzerathsche Gesetz. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 659–688.

Grzybek, P., Stadlober, E. (2007). Do we have problems with the Arens' law? A new look at the sentence-word relation. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text*. Berlin, New York: de Gruyter, 205–217.

Meyer, P. (2007). Two semi-mathematical asides on Menzerath-Altmann´s law. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text.* Berlin, New York: de Gruyter, 448–460.

8.8 NARANAN-BALASUBRAHMANYANOVO ROZDĚLENÍ

Problém

Aplikujte Naranan-Balasubrahmanyanovo rozdělení na rankovou frekvenční distribuci fonémů/písmen. Aplikujte ji také na rankovou frekvenční distribuci slov. Prokažte, že tento model může být odvozen z tzv. sjednocené teorie (Unified Theory).

Postup

Naranan-Balasubrahmanyanovo rozdělení je definováno jako $P_x = Ce^{-a/x_x-b}$, $x = 1, 2, 3,...$ kde a, b jsou parametry a C je normalizační konstanta. Vytvořte větší výběrový soubor fonémů nebo písmen (grafémů) z textu (nebo z dostupné literatury) a pokuste se aplikovat uvedené rozdělení na empirickou rankovou frekvenční distribuci. Odvoďte několik jednoduchých odhadů parametrů a testujte shodu modelu s daty pomocí chí-kvadrát kritéria.

Literatura

Krylov, J. K. (1987). Stacionarnaja model´ porožděnija svjaznogo teksta. *Acta et Commentationes Universitatis Tartuensis 774*, 81–102.

Balasubrahmanyan, V. K., Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics 3(3)*, 177–228.

Naranan, S., Balasubrahmanyan, V. K. (2007). Statistical analogs in DNA sequences and Tamil language texts: rank frequency distribution of symbols and their application to evolutionary genetics and historical linguistics. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text*. Berlin, New York: de Gruyter, 484–497.

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics 3(1)*, 38–50.

Wimmer, G., Altmann, G. (2006). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the science of language. Word length studies and related issues*. Boston: Kluwer, 93–117.

8.9 ORDOVO KRITÉRIUM

Problém

Vezměte skupinu textů, spočítejte distribuce nějaké proměnné a pokuste se ji charakterizovat pomocí Ordova kritéria.

Postup

Pokud jste spočítali frekvence, vypočítejte první tři momenty:

$$m_1' = \frac{1}{N} \sum_{x=x_{min}}^{x_{max}} x f_x \ , \qquad \text{(průměr)}$$

$$m_2 = \frac{1}{N} \sum_{x=x_{min}}^{x_{max}} (x - m_1')^2 f x \ , \quad \text{(variance, druhý centrální moment)}$$

$$m_3 = \frac{1}{N} \sum_{x=x_{min}}^{x_{max}} (x - m_1')^3 f x \ , \quad \text{(asymetrie, třetí centrální moment)}$$

a stanovte indikátory:

$$I = \frac{m_2}{m_1'}\,, \qquad\qquad S = \frac{m_3}{m_2'}\,.$$

Potom zaneste body <*I*, *S*> jednotlivých textů do kartézské soustavy souřadnic, kde můžete vidět pozice a vzdálenosti mezi texty.

Tento graf je typem jednoduché klasifikace, která umožňuje vytvářet hypotézy o stavu a vývoji dané vlastnosti. Pokuste se analyzovat *délku věty* v různých žánrech a ve stejném žánru v různých obdobích.

Rozložte dvojčleny v centrálních momentech a zjednodušte je. Analyzujte 10 poetických a 10 vědeckých textů ve vašem jazyce. Pro každý text vypočítejte distribuci délky vět. Potom použijte Ordovo kritérium k zobrazení rozdílu mezi těmito dvěma žánry.

Pokud chcete porovnávat distribuce frekvencí slov v různých jazycích, testujte, zda leží na různých přímkách.

Literatura

Best, K. H. (2005). Wortlänge. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 260–273.

Oakes, M. P. (2007). Ord´s criterion with word length spectra for the discrimination of texts, music and computer programs. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text*. Berlin, New York: de Gruyter, 508–519.

Ord, J. K. (1972). *Families of frequency distributions*. London: Griffin.

Popescu, I.-I., Vidya, M. N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008). *Word frequency studies*. Berlin, New York: de Gruyter.

8.10 INDEX OPAKOVÁNÍ A ENTROPIE

Problém

Index opakování je definován jako

$$R = \sum_x p_x^2 = \frac{1}{N^2} \sum f_x^2 \ ,$$

entropie je definována jako

$$H = -\sum_x p_x \ ld \ p_x = ld \ N - \frac{1}{N} \sum_x f_x \ ld \ f_x \ ,$$

kde N je velikost výběrového souboru, p_x je pravděpodobnost dané entity, f_x je absolutní frekvence této entity a *ld* je logaritmus o základu 2. Ověřte, zda jsou tyto indexy závislé na velikosti inventáře.

Postup

Spočítejte fonémy (v různých jazycích) a slova. Vytvořte rankovou frekvenční distribuci slov a jejich distribuční spektrum. Vypočítejte oba výše uvedené indexy a zkoumejte jejich vztah k velikosti inventáře fonémů a k velikosti slovníku jednotlivých textů. Pokuste se najít závislost.

Literatura

Altmann, G., Lehfeldt, W. (1980). *Einführung in die quantitative Phonologie*. Bochum: Brockmeyer.

Popescu, I.-I., Vidya, M. N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008). *Word frequency studies*. Berlin, New York: de Gruyter.

8.11 VELIKOST VÝBĚROVÉHO SOUBORU

Problém

Velikost výběrového souboru je základním problémem v každém kvantitativně lingvistickém výzkumu. Pro aplikaci jakéhokoliv testu je nutné mít dostatečný počet případů zkoumaného jevu. Frekvenční distribuce musí mít dostatečný počet případů v každé x-třídě. Body na frekvenční křivce jsou obvykle průměry dostatečně velkého množství případů. Pro frekvence fonémů se používá Kubáčkova metoda. Pokuste se vynalézt induktivní metodu, pomocí níž zjistíte dostatečnou velikost výběrového souboru pro jakoukoliv jazykovou jednotku.

Postup

Jako příklad nám může posloužit metoda pro frekvence fonémů. Pokud nevíte, kolik fonémů musí být vybráno, vyzkoušejte následující empirickou metodu. Z nějakého textu vyberte 1 000 fonémů (písmen) a zapište jejich frekvence do sloupce. Vyberte dalších 1 000 fonémů a jejich frekvence přidejte předchozím do nového sloupce. Tento postup opakujte, dokud nebudete mít spočítáno 10 000 fonémů. Potom spočítejte relativní frekvence fonémů v každém sloupci tabulky. Sečtěte absolutní rozdíly mezi každým párem sousedních sloupců a pozorujte pokles tohoto součtu. Aplikujte funkci $y = a10^{-bx}$ (iterativně) k této klesající řadě, kde x je pořadové číslo sloupce. Najděte bod x, ve kterém $y = \delta$, přičemž $\delta = 1/10K$ (K = počet fonémů v inventáři). Nezbytná velikost výběrového souboru je při daném x

$$N = 1000(x + 1).$$

Zobecněte tuto metodu a použijte ji pro zjištění dostatečné velikosti výběrového souboru slabik, morfů, a dokonce i slov. Porovnejte výsledky vašich výpočtů s klasickými statistickými metodami.

Literatura

Altmann, G., Lehfeldt, W. (1980). *Einführung in die quantitative Phonologie*. Bochum: Brockmeyer.

Kubáček, L. (1994). Confidence limits for proportions of linguistic entities. *Journal of Quantitative Linguistics 1(1)*, 56–61.

8.12 PROBLÉM NEKONEČNA

Problém

Pokuste se vyřešit nebo alespoň diskutovat problém nekonečna v jazyce. Některé matematické modely ukazují, že rozsah určitých jazykových jednotek může být nekonečný, např. počet vět v jazyce, velikost slovní zásoby, délka věty, dokonce i délka slova atd. Víme, že v jazyce existují určité mezní hodnoty, např. počet fonémů, počet slabik, počet slov uložených v paměti (vezměte v úvahu také polygloty a specialisty!) – existuje však nějaký limit pro počet významů? Pokud ano, zkuste najít opodstatnění. Pokud ne, vysvětlete tento fakt.

Postup

Začněte se Zipfovými unifikačními a diverzifikačními silami a zvažte důsledky hypotetické existence slova s nekonečným množstvím významů. V případě délky vět vezměte v úvahu Köhlerovu interpretaci Menzerathova zákona. V případě inventáře fonémů zvažte problém minimální distinktivnosti. V případě počtu různých slabik začněte s fenoménem nezbytnosti

adekvátní redundance atd. Zobecněte problém a vezměte přitom v úvahu köhlerovské „požadavky".

Literatura

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics. An international handbook*. Berlin, New York: de Gruyter, 760–774.

Köhler, R. (1989). Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus. Chapter 7. In: Altmann, G., Schwibbe, M. H., *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms, 108–112.

Zipf, G. K. (1935). *The psycho-biology of language. An introduction to dynamic philology*. Boston: Houghton Mifflin.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Reading, Mass: Addison-Wesley.

8.13 TĚSNOST/KOHEZE

Hypotézy

„…čím častěji se dva prvky vyskytují v řadě za sebou, tím pevnější bude jejich konstituentní struktura." (Bybee, Hopper 2001: 14; Bybee, Scheibman 1999). Jednoduchým příkladem jsou případy, kde se spojila dvě slova kvůli jejich častému spoluvýskytu a nyní se chovají jako slovo jedno, např. *want to > wanna; going to > gonna; I am > I´m; can not > can´t; do not > don´t; I don´t know > I dunno; would have >would´ve.*

„U dvojic slov, která jsou často používána společně, ať už je jejich lexikální nebo gramatický význam jakýkoliv (*don´t you*, *told you*, *that you*, *last year*), se častěji projevuje efekt koartikulace než u slov, která nejsou společně používána tak často." (Bybee, Hopper 2001: 7).

Postup

Pokuste se navrhnout způsob měření těsnosti/koheze konstituentů. Potom testujte dvojice slov, které se v korpusu vyskytují společně signifikantně často. Zkuste potvrdit hypotézu *těsnost = f(frekvence spoluvýskytu)*. Hypotézu testujte na 100 párech slov. Následně se pokuste dokázat, že čím častěji se nějaký pár vyskytuje, tím větší je jeho těsnost, tj. *koheze = f(frekvence spoluvýskytu).*

Literatura

Boyland, J. T. (1996). *Morphosyntactic change in progress: a psycholinguistic approach*. Diss: Linguistics Department, University of California.

Bybee, J. (2000). Lexicalization of sound change and alternating environment. In: Broe, M., Pierrehumbert, J. (eds.), *Laboratory V: Language acquisition and the lexicon*. Cambridge: Cambridge University Press, 250–268.

Bybee, J., Hopper, P. (2001). Introduction to frequency and the emergence of linguistic structure. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 1–24.

Bybee, J., Scheibman, J. (1999). The effect of usage on degree of constituency: the reduction of don´t in American English. *Linguistics 37*, 575–596.

Fan, F., Altmann, G. (2007). Measuring the cohesion of compounds. In: Kaliučenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology*. Černivci: RUTA, 177–189.

Krug, M. (1998). String frequency: a cognitive motivating factor in coalescence, language processing and linguistic change. *Journal of English Linguistics* 26, 286–320.

Krug, M. (2001). Frequency, iconicity, categorization: evidence from emerging modals. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 310–335.

8.14 ZIPFŮV A ZIPFŮV-MANDELBROTŮV ZÁKON

Problém

Projděte historii Zipfova a Zipfova-Mandelbrotova zákona.

Postup

Na začátku shromážděte patřičnou literaturu. Připravte pokud možno kompletní bibliografii týkající se těchto zákonů. Nevynechejte ruskou literaturu. Potom projděte všechny vzorce, které byly vytvořeny v souvislosti s těmito zákony. Dále sledujte všechny způsoby odvození jednotlivých vzorců. Pokuste se je rozdělit do několika skupin podle různého teoretického pozadí. Rozlišujte lingvistické argumentace od argumentací obecných, ale prezentujte je všechny. Sledujte, jak se tento zákon používá v jiných vědních oborech. Zipfův zákon je nejznámější mocninný zákon, Mandelbrotova verze je jeho zobecněním. Představte všechna zobecnění, která najdete v literatuře. Poukažte také na odlišné případy.

Literatura

Glottometrics 3–5, 2002. [a collection of articles to honor G. K. Zipf].

Guiter, H., Arapov, M. V. (eds.) (1982). *Studies on Zipf´s law*. Bochum: Brockmeyer.

9 Výzkumné projekty

9.1 FRUMKINOVÉ ZÁKON (VÝSKYT SLOV V DANÝCH PASÁŽÍCH TEXTU)

Problém

Rozdělte dlouhý text na pasáže o stejné délce n (např. 50, 100, 150, 200,… slov). Potom vyberte slovo (ne příliš málo frekventované) a spočítejte jeho výskyt v jednotlivých pasážích, tj. spočítejte počet pasáží, ve kterých se dané slovo vyskytuje $x = 0, 1, 2\ldots$ krát. Distribuce výskytů slova v pasážích bude mít podobu negativního hypergeometrického rozdělení nebo jeho limitních případů (binomické, negativní binomické, geometrické, Poissonovo rozdělení).

Postup

Použijte FITTER nebo jiný vhodný software pro nalezení vhodného rozdělení. Tento problém má více aspektů:

(1) Pokud délka pasáží vzrůstá, mění se parametry distribuce nebo distribuce konverguje k nějaké limitní formě. Zkoumejte problém na několika různých slovech. Analyzujte všechny slovní druhy a zjistěte, zda má délka pasáží vliv na parametr nebo zda se různé (limitní) změny objevují v různých slovních druzích.

(2) Stanovte stejnou délku pasáže a vypočítejte distribuce mnoha slov jednoho slovního druhu. Zjistěte frekvenci daného slova v celém textu a pokuste se naléztvztah mezi (relativní) frekvencí slova a jedním z parametrů distribuce.

(3) Vyvoďte závěry z formy distribuce (nebo z hodnot parametrů) týkající se slovního druhu, ke kterému by dané slovo mělo patřit.

(4) Uveďte empirické podmínky, za kterých základní negativní hypergeometrické rozdělení konverguje ke svým limitním případům (která slova, které slovní druhy, jaká frekvence, jaká délka slov atd.).

(5) Proveďte následující:

   (a) Vypracujte závěry o sémantice slova v textu na základě druhu rozdělení (nebo jeho parametrů).

   (b) Vyvoďte závěry o psychickém/emocionálním stavu daného autora na základě odchylek distribuce stejných slov.

   (c) Pokuste se nalézt rozdíly mezi jazyky, žánry, styly, autory na základě distribuce výskytů slov v pasážích.

Tento problém je téma pro výzkumný projekt sestávající z týmu lingvistů, psychologů a programátorů.

Literatura

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Altmann, G., Burdinski, V. (1982). Towards a law of word repetitions in textblocks. *Glottometrika 4*, 147–167.

Bektaev, K. B., Lukjanenkov, K. F. (1971). O zakonach raspredelenija edinic pismennoj reči. In: Piotrowski, R. G. (ed.), *Statistika reči i avtomatičeskij analiz teksta*. Leningrad: Nauka, 47–112.

Best, K.-H. (2001, 2003). *Quantitative Linguistik. Eine Annäherung*. 2. überarbeitete und erweiterte Auflage. Göttingen: Peust & Gutschmidt.

Best, K.-H. (2005). Sprachliche Einheiten in Textblöcken. *Glottometrics 9*, 1–12.

Billmeier, G. (1968). Über die Signifikanz von Auswahltexten. Untersuchung auf der Grundlage von Zeitungstexten. In: Moser, H. u. a. (Hrsg.), *Forschungsberichte des Instituts für deutsche Sprache 2*, 126–171.

Brainerd, B. (1972a). Article use as an indirect indicator of style among Englishlanguage authors. In: Jäger, S. (ed.), *Linguistik und Statistik*. Braunschweig: Vieweg, 11–32.

Frumkina, R. M. (1962). O zakonach raspredelenija slov i klassov slov. In: Mološnaja, T. N. (ed.), *Strukturno-tipologičeskie issledovanija*. Moskva: ANSSSR, 124–133.

Herdan, G. (1956). *Language as choice and chance*. Groningen: Nordhoff.

Knauer, K. (1955). Grundfragen einer mathematischen Stilistik. *Forschungen und Fortschritte 29*, 140–149.

Köhler, R. (2001). The distribution of some syntactic construction types in text blocks. In: Uhlířova, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of L. Hřebíček*. Trier: Wissenschaftlicher Verlag, 136–148.

Leopold, E. (1998). *Stochastische Modellierung lexikalischer Evolutionsprozesse*. Hamburg: Kovač.

Maškina, L. E. (1968). *O statističeskich metodach issledovanija leksiko-grammatičeskoj distribucii.* Minsk, Diss.

Morton, A. Q., Levison, M. (1966). Some indicators of authorship in Greek prose. In: Leed, J. (ed.), *The computer and literary style*. Kent, Ohio: Kent State UP, 141–179.

Mosteller, F., Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist*. Reading, Mass, Addison-Wesley.

Muller, Ch. (1972). *Einführung in die Sprachstatistik*. München: Hueber.

Paškovskij, V. E., Srebrjanskaja, I. I. (1971). Statističeskie ocenki pis'mennoj reči bol'nych šizofreniej. *Inženernaja lingvistika*. Leningrad.

Piotrowski, R. G. (1984). *Text – Computer – Mensch*. Bochum: Brockmeyer.

Piotrowski, R. G., Bektaev, K. B., Piotrowskaja, A. A. (1985). *Mathematische Linguistik*. Bochum: Brockmeyer.

Suhren, S. (2002). *Untersuchung zum Gesetz von Zwirner, Zwirner und Frumkina am Beispiel des niederdeutschen „De lütte Prinz“*. Staatsexamensarbeit, Göttingen.

Zwirner, E., Ezawa, K. (Hrsg.) (1966, 1968, 1969). *Phonometrie, Erster-Dritter Teil*. Basel, New York: Karger.

Zwirner, E., Zwirner, K. (1935). Lauthäufigkeit und Zufallsgesetz. *Forschungen und Fortschritte* 11, Nr. 4, 43–45. [Also in: Zwirner & Ezawa (Hrsg.), Dritter Teil, 55–59].

Zwirner, E., Zwirner, K. (1938). Lauthäufigkeit und Sprachvergleichung. *Monatsschrift für höhere Schulen* 37, 246–253. [Also in: Zwirner & Ezawa (Hrsg.), Dritter Teil, 68–74].

9.2 SKALIČKŮV TYPOLOGICKÝ SYSTÉM

Hypotéza

Skaličkův systém zahrnuje následující vlastnosti jazyka:

(1) Délka kořenu.

(2) Délka slova.

(3) Rozlišování slovních druhů.

(4) Komplexita slova.

(5) Konverze.

(6) Počet afixů.

(7) Rozsah derivace.

(8) Počet synsémantik.

(9) Délka afixu.

(10) Tvoření kompozit.

(11) Počet prepozic a postpozic.

(12) Homonymie afixů.

(13) Synonymie afixů.

(14) Pevnost slovosledu.

(15) Flexe.

(16) Vnitřní flexe.

(17) Počet klauzí.

(18) Počet koncovek ve slově.

(19) Morfémová diskontinuita.

(20) Existence infinitivů, participií, slovesných substantiv.

(21) Počet vokálů a konsonantů.

(22) Rozsah shody.

(23) Rozlišování kořene a afixů.

(24) Rozlišování flexe a derivace.

(25) Příznakovost věty.

(26) Vokální harmonie.

(27) Supletivismus.

(28) Tvoření členů.

(29) Posesivita.

(30) Rozsah deklinace.

Všechny tyto vlastnosti spolu souvisejí. Pokuste se potvrdit existenci vzájemných vztahů, a to jak empiricky, tak teoreticky.

Postup

Kvantifikujte a měřte alespoň některé z uvedených vlastností v textech nebo v korpusu. U vlastností s mnoha hodnotami postačí analyzovat pro částečné potvrzení hypotézy jeden jazyk, v případě binárních rysů je však nezbytné analyzovat alespoň 10 různých jazyků. Je nezbytné přečíst některé Skaličkovy práce, např. Skalička (1966). Nezapomeňte, že před kvantifikací, měřením a testováním je nezbytné jasně definovat hypotézu. Začněte s dvěma libovolnými vlastnostmi. Pokud řešíte problém zahrnující více než dvě vlastnosti, vytvořte vztahový diagram a postupně jej rozšiřujte (viz Köhler 1986). Velmi rozsáhlý seznam vlastností ve Skaličkových sebraných spisech (2005–2006) může být použit k dohledání jeho článků napsaných v němčině, angličtině, francouzštině, ruštině nebo maďarštině.

Literatura

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics* 26, 178–194.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Skalička, V. (1964a), Konsonantenkombinationen und linguistische Typologie. *Travaux linguistiques de Prague 1,* 111–114.

Skalička, V. (1964b). Typologie a konfrontační lingvistika. *Československá rusistika 7,* 210–212.

Skalička, V. (1966). Ein „typologisches Konstrukt". *Travaux linguistiques de Prague* 2, 157–163.

Skalička, V. (2005–2006). *Souborné dílo I–III.* Praha: Nakladatelství Karolinum.

9.3 SYNONYMIE

Problém

Synonymie je součástí několika kontrolních cyklů, jejichž ostatními elementy jsou různé vlastnosti slov. Každý bod níže referuje k potenciální vlastnosti slova, např. potencionální hypotézu týkající se synonymie. Jde o tyto vlastnosti:

(1) Délka měřená v počtu fonémů, slabik, morfémů, mor apod.

(2) Frekvence slova v korpusu.

(3) Polysémie jako počet významů ve slovníku (nebo ve Wordnetu).

(4) Polytextualita jako počet textů nebo kolokací slova (kontextů), viz také bod 11.

(5) Morfologický statut: jednoduchý, odvozený, reduplikovaný, kompozitní.

(6) Atribuce slovních druhů: čím větší je počet slovních druhů, ke kterým slovo náleží, tím více má synonym (srov. přímou konverzi ve Wordnetu).

(7) Morfologická produktivita: čím více je možných odvozenin, kompozit nebo reduplikací daného slova, tím více synonym toto slovo má (databáze pro němčinu je na internetu).

(8) Stáří slova, určeno v počtu století od prvního výskytu slova v jeho v psané podobě. Čím je slovo starší, tím více má synonym (záleží na slovním druhu). Těžko ověřitelné.

(9) Původ: počet historických přejímek slov, např. z latiny do francouzštiny a ruštiny, z arabštiny přes angličtinu do němčiny atd. Čím delší je „cesta“ daného slova, tím více má synonym.

(10) Valence sloves: počet aktantů, se kterými může být sloveso spojeno. Valence zvyšuje polysémii, což zvyšuje synonymii.

(11) Speciální případy v některých jazycích: počet předložek, se kterými sloveso tvoří frázi (*get in*, *get out*, *get around*, *get off*, *get out of*, *get from under*, *get through*,…). Čím více je předložkových frází, tím více je synonym, protože mnoho takovýchto frází může být nahrazeno jedním slovem.

(12) Počet gramatických kategorií slova (pád, číslo, čas,…). Čím více kategorií, tím větší je synonymie. Kategorie umožňují, aby se slovo mohlo vyskytnout v různých kontextech, tudíž platí, že rostoucí polytextualita → rostoucí polysémie.

(13) Emocionalita vs. pojmovost (např. *matka* vs. *banka*). Musí být realizováno prostřednictvím testu probandů a musí být navržena škála. Hypotéza není známa (může platit v obou směrech, ale předpokládá se, že čím větší je emocionalita, tím více je synonym, např. *Ty svině!*).

(14) Pollyannin princip: pozice slova na škále dobré–špatné. (Testujte probandy.)

(15) Abstraktnost vs. konkrétnost, např. *krása* vs. *revolver*. (Musí být vytvořen speciální škálovací postup).

(16) Určitost vs. obecnost (např. *revolver* vs. *nástroj*).

(17) Dogmatismus slova (např. *muset* vs. *moci*, *vše* vs. *něco*, *vždy* vs. *někdy*).

(18) Počet asociací (konotativní síla). Použijte slovníky asociací. Čím více asociací, tím více synonym.

(19) Počet možných funkcí, které může mít slovo ve větě (např. slovo může být subjektem, predikátem, objektem, komplementem,…).

(20) Diatopická variantnost slova: čím více forem se realizuje v dialektech, tím více synonym je tvořeno. (Může být měřeno jako počet tvarů v jazykovém atlasu.)

(21) Diskurzivní vlastnosti slova: indikuje slovo asociaci se sociální třídou?

(22) Stupeň standartizace (vyšší styl, neutrální styl, městská mluva, slang…).

(23) Diverzifikace: Do kolika slovních druhů může být slovo transformováno prostřednictvím afixů (nikoliv konverzí!), např. v němčině: *Bild* (substantivum), *bildhaft* (adjektivum/adverbium), *bilden* (sloveso).

(24) Původ: (a) původní slovo, (b) kalk, (c) výpůjčka.

(25) Počet neměnných frází určitých slovních tvarů (speciální případ polytextuality, řekněme polytextuality I).

Každá z těchto vlastností podporuje nebo omezuje tvoření synonym (případně musí být vyřazena, pokud se chová vzhledem k synonymii neutrálně). Měly by být vytvořeny a následně testovány další hypotézy. Musí být postupně tvořeny jednotlivé cykly a nakonec vytvořen synergetický kontrolní cyklus.

Postup

Vyberte náhodně 500 slov ze slovníku synonym a pro každé slovo spočítejte počet jeho synonym. Potom zkoumejte jednu z výše uvedených vlastností a zkuste dokázat, že $synonymie = f(vlastnost)$. Testujte krok za krokem všechny hypotézy, případu od případu budete muset vymyslet způsob měření jednotlivých vlastností. Pokud se vyskytne závislost, zobrazte ji v diagramu, v němž spojíte synonymii s danou vlastností. Pokračujte, dokud

nebudou ověřeny vztahy všech výše uvedených vlastností se synonymií. Následně se pokuste nalézt i závislosti mezi jednotlivými vlastnostmi.

Pokračování

Synonymie může vzniknout několika způsoby:

(1) Za určitých okolností dané slovo nedovoluje vyjádřit potřebný význam a je nahrazeno jiným slovem (např. ironie, postoj, slang,... např. v latině *caput, testa*).

(2) Takové okolnosti mohou být reprezentovány určitým kontextem, ve kterém se vyskytují (takovým kontextem je např. zbytek věty). Tento případ spojuje synonymii s polytextualitou. Každý kontext nepatrně mění význam slova. K vyjádření hledaného významu se použije vhodnější slovo.

(3) Každé slovo má tendenci k rostoucí polysémii, ale některé významy se se slovem přestanou pojit a jsou vyjádřeny jinými slovy kvůli potřebě specifikace.

(4) Najděte další příčiny růstu synonymie.

Literatura

Popescu, I.-I., Vidya, M. N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008). *Word frequency studies*. Berlin, New York: de Gruyter.

Wimmer, G., Altmann, G. (2001). Two hypotheses on synonymy. In: Ondrejovič, S., Považaj, M. (eds.), *Lexicographica´99. Zborník na počest Kláry Buzássyovej*. Bratislava: Veda, 218–225.

Ziegler, A. (2001). Zum Gesetz der Synonymie. Modellanpassungen im Deutschen und Englischen. In: Ondrejovič, S., Považaj, M. (Hrsg.), *Lexicographica´99. Zborník na počest Kláry Buzássyovej*. Bratislava: Veda, 230–236.

Ziegler, A., Altmann, G. (2001). Beziehung zwischen Synonymie und Polysemie. In: Ondrejovič, S., Považaj, M. (Hrsg.), *Lexicographica 99. Zborník na počest Kláry Buzássyovej*. Bratislava: Veda, 226–229.

9.4 FREKVENCE SLOV A PŘÍBUZNÉ VLASTNOSTI

Hypotéza

Mnoho vlastností slov je spojeno s jejich frekvencí.

Postup

Tato hypotéza je velmi široká a může být testována jen postupně. Příbuzné vlastnosti jsou takové, které mohou být stanoveny pro jednu a tutéž jednotku, např. slovo. Následně jsou všechny sledované vlastnosti operacionalizovány (kvantifikovány). Ze slovníku jsou utvořeny náhodné výběry a jsou měřeny jejich vlastnosti. Pak je možné zkoumat vztah frekvence vybraných slov vzhledem k daným vlastnostem.

Může být také proveden „obrácený" způsob: nejdříve spočítejte frekvence všech lemmat v textu, potom sledujte některé vlastnosti těchto lemmat a testujte jejich vztah k frekvenci.

Některé vlastnosti mohou být měřeny pouze na nominální škále. Pokuste se nicméně najít způsob, jak vytvořit ordinální škálu. Pro usnadnění řešení tohoto problému použijte seznam 27 vlastností slov z Popescu et al. (2008). Tento seznam rozhodně není kompletní, mohou být proto přidány další vlastnosti (srov. seznam vlastností v části 9.3, „Synonymie").

(1) Délka: měřena v počtu fonémů, slabik, morfémů, mor nebo morfémů apod. Někdy je tato vlastnost označována jako materiální komplexita.

(2) Polysémie: počet významů ve slovníku.

(3) Morfologický status: jednoduché, reduplikované, odvozené, složené slovo.

(4) Počet slovních druhů, ke kterým dané slovo náleží, např. konverzí (*the hand, to hand*).

(5) Polytextualita: počet textů, ve kterých se objevuje, nebo počet kontextů (kolokace).

(6) Produktivita: počet odvozených slov, kompozit, reduplikací, které mohou být utvořeny daným slovem. Data můžete nalézt na internetu.

(7) Stáří: počet let nebo století od prvního dokladu slova v textu.

(8) Původ: kolika jazyky dané slovo prošlo než se stalo součástí analyzovaného jazyka.

(9) Valence u sloves: počet různých pádů a předložek, se kterými se může vyskytnout.

(10) Počet gramatických kategorií slova: pád, číslo, rod, čas, osoba, způsob atd. nebo počet afixů, se kterými se může kombinovat (např. ne všechna slovesa mohou být kombinována se všemi prefixy).

(11) Stupeň emocionality vs. pojmovosti. Srovnejte například emocionalitu slov *matka* a *tužka*.

(12) Pollyannin princip: stupeň slova na škále „dobré–špatné".

(13) Stupeň abstraktnosti vs. konkrétnosti slova, např. *krása* vs. *tužka*.

(14) Určitost vs. obecnost, např. *tužka* vs. *nástroj*.

(15) Stupeň dogmatismu, např. *muset* vs. *moci*; *vše* vs. *něco*; *vždy* vs. *někdy*.

(16) Počet asociací (= konotativní potenciál), jež vznikají když je slovo vnímáno sluchem či zrakem. Existují slovníky asociací.

(17) Synonymie: počet synonym ve slovníku.

(18) Počet možných funkcí, které může mít slovo ve větě (např. slovo může být subjektem, predikátem,...).

(19) Diatopická variantnost: v kolika místech může být slovo nalezeno v jazykovém atlasu.

(20) Nářeční konkurence: kolik variant daného slova se vyskytuje v jazykovém atlasu.

(21) Diskurzivní vlastnosti slova: v jakém rozsahu slovo indikuje atribuci k sociální třídě?

(22) Stupeň standartizace: spisovný jazyk, sociolekt, argot atd.

(23) Diverzita: do kolika slovních druhů může dané slovo patřit, pokud jsou z něj vytvořeny odvozeniny, např. v němčině *Bild* (N) -> *bildhaft* (Adj), *bilden* (V), *bildlich* (Adj, Adv).

(24) Původ: původní slovo, výpůjčka, kalk, lidová etymologie atd.

(25) Frazeologie: v kolika idiomech se dané slovo nachází?

(26) Stupeň slovesné aktivity, např. *spát* vs. *běžet*.

(27) Stupeň vyjádření vlastnosti u adjektiv, např. *pěkný*, *hezký*, *nádherný*.

Snažte se přidat další vlastnosti a hledejte jejich vztahy s frekvencí.

Literatura

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Popescu, I.-I., Vidya, M. N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008). *Word frequency studies*. Berlin, New York: de Gruyter.

KATALOGIZACE V KNIZE - NÁRODNÍ KNIHOVNA ČR

Strauss, Udo
Kvantitativní lingvistika : vybrané problémy 1 / Udo Strauss, Fengxiang Fan, Gabriel Altmann ; [překlad Miroslav Kubát, Radek Čech].
-- Olomouc : Univerzita Palackého v Olomouci, 2014. -- 196 s. -- (Qfwfq ; sv. 31)
Název originálu: Problems in quantitative linguistics 1
Přeloženo z angličtiny
ISBN 978-80-244-4350-8

81'324
- kvantitativní lingvistika
- monografie

81 - Lingvistika. Jazyky [11]

Kvantitativní lingvistika. Vybrané problémy 1
Udo Strauss, Fengxiang Fan, Gabriel Altmann

31. svazek Edice Qfwfq

Výkonný redaktor: Agnes Hausknotzová
Odpovědná redaktorka VUP: Jana Kreiselová
Jazyková redakce: Radek Čech
Sazba a obálka: Martina Šviráková

Vydala a vytiskla Univerzita Palackého v Olomouci
Křížkovského 8, 771 47 Olomouc
www.upol.cz/vup
e-mail: vup@upol.cz
Olomouc, 2014
1. vydání, 196 stran
č. z. 2014/946

ISBN 978-80-244-4350-8