Reinhard Köhler
Gabriel Altmann

Edice Qfwfq

Kvantitativní lingvistika

Vybrané problémy 2

Olomouc
2014

Překlad:
Jiří Bareš, Radek Čech

Recezní posudek překladu:
Prof. RNDr. Ing. Lubomír Kubáček, DrSc., dr. h. c.

Přeloženo podle:
Reinhard Köhler, Gabriel Altmann: Problems in quantitative linguistics 2.

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 978-80-244-4324-9

Obsah

Předmluva

„Nejenže na počátku každého výzkumu stojí nějaký problém: výzkum spočívá v neustálém řešení problémů."

(Bunge, M. [2007]. *Philosophy of Science, Vol. 1: From problem to theory.* New Brunswick, London: Transaction Publishers, 187.)

Každý mladý vědec si nejprve musí najít nějaký vědecký problém. Dalším úkolem je pak jeho vyřešení. Nalezeným řešením však problém nekončí. S každým řešením naopak vyvstává řada nových problémů. V každém vědním oboru by tak bylo čas od času dobré provést rekapitulaci aktuálních problémů, upozornit na některé nové otázky a současně poukázat na další aspekty starých problémů.

V této publikaci představujeme soubor problémů v oblasti kvantitativní lingvistiky, alespoň do té míry, pokud je možné vysledovat „ariadninu nit" ve změti nerovnoměrně rozvinutých dílčích disciplín tohoto předmětu vědeckého zkoumání. Z důvodu značné roztříštěnosti, kterou se celý tento obor vyznačuje, se snažíme nenásilně upozornit čtenáře na možnosti jeho sjednocení, jež se může stát východiskem pro rozvíjení další teorie. V dnešní době si lze jen obtížně představit, že by v rámci empirického vědního oboru mohla nějaká teorie vzniknout bez alespoň elementární kvantifikace. Přestože problémy, které zde prezentujeme, stále obnášejí i množství kvalitativní práce, snažíme se čtenáře přesvědčit, aby danou problematiku vnímali z kvantitativního hlediska, aby usilovali o elementární kvantitativní řešení, hledali souvislosti mezi některými problémy a již existujícími teoriemi nebo poukazovali na nové oblasti zkoumání.

V prvním svazku této řady představili autoři některé problémy z oblasti fonologie, písma, gramatiky, lexikologie, textologie, sémantiky, synergetiky, psycholingvistiky, typologie, dále různé obecné problémy a také problematiku délky a frekvence ve vztahu k jiným vlastnostem. Také v tomto svazku se zabýváme většinou z výše uvedených oblastí, avšak navíc je zde pojednáno o řadě problémů, které souvisejí například s pragmatikou, příslovími, dramatem, filozofií vědy, motivy nebo dialektologií. Pokud se čtenář rozhodne některý z daných

problémů řešit, doporučujeme, aby si nejprve prostudoval první svazek této řady, kde se může seznámit s podobnými problémy v elementárnější podobě. Pokud se nám podaří nějaký problém úspěšně vyřešit, měli bychom se jej snažit zobecnit, ověřit daný výsledek na datech z několika jazyků nebo textů, hledat odchylky, nepravidelnosti v chování sledovaných jevů či jednotek, doplnit jej o další podmínky a systematizovat jej, tj. ukotvit v obecnějším rámci, z něhož jej lze odvodit.

Pokud narazíme na „zapeklitější" problémy, může se stát, že v první fázi budeme muset postupovat čistě induktivně, např. tak, že na zkoumaná data mechanicky aplikujeme nějakou jednoduchou funkci, avšak v dalším kroku by již zkusmo ověřovaná funkce měla obstát v konfrontaci s otázkou „proč zvolit právě tuto funkci?", což má k budoucímu vysvětlení blíže než pouhý slovní popis odhaleného jevu.

Problémy prezentované v této publikaci mají různou povahu: najdeme zde zadání odpovídající školním cvičením z kvantitativní lingvistiky, ale také výchozí materiál pro publikační činnost či témata vhodná k dalšímu zpracování v rámci výzkumných projektů.

Rádi bychom čtenáře vyzvali, aby editora časopisu *Journal of Quantitative Linguistics* [Dostupné z: http://www.ldv.uni-trier.de/index.php?koehler] a editora časopisu *Glottometrics* [Dostupné z: www.gabrielaltmann.de] informovali o publikační činnosti stavějící na některém z problémů nastíněných ať už v tomto nebo prvním svazku. Řešení jednotlivých problémů mohou být rovněž předložena k publikování v některém z těchto časopisů.

Dále bychom také čtenáře chtěli vyzvat, aby na výše uvedené adresy zasílali příspěvky, v nichž by poukazovali na případné další nové problémy.

R. K., G. A.

1 Fonologie a písmo

1.1 ZIPFOVA ASIMILACE

Hypotéza

„…každá asimilace vypovídá o oslabení či nestabilitě asimilované hlásky, přičemž toto oslabení či nestabilitu primárně způsobuje nadměrná relativní frekvence asimilované hlásky." (Zipf 1935/68: 109). Ověřte tuto hypotézu.

Postup

Pořiďte si výčet všech fonologických asimilací v analyzovaném jazyce. Čerpat můžete z nějaké učebnice fonologie (případně z vlastních znalostí) a z přehledu relativních frekvencí hlásek v daném jazyce. Které z následujících tvrzení odpovídá výsledku, ke kterému jste dospěli?

(a) Hypotéza je pravdivá,

(b) naopak, asimilované hlásky mají zřejmě spíše nižší frekvenci,

(c) hláska, jež vyvolala asimilaci, má velmi vysokou frekvenci,

(d) obě hlásky (asimilovaná i asimilující) jsou relativně vzácné.

Pokud neplatí (a), hypotézu zobecněte a ověřte ji na několika jazycích.

Můžete také případně zjistit, za jakých podmínek hypotéza v daném jazyce platí, a na základě těchto podmínek ji modifikovat.

Pokud je hypotéza pravdivá, formulujte alespoň empirický vzorec, kterým je možné tento vztah vyjádřit. Proveďte systematizaci dané hypotézy tím, že ji zasadíte do určitého řídicího cyklu nebo prokážete, že je důsledkem obecnějších mechanismů.

Literatura

Zipf, G. K. (1935/1968). *The psycho-biology of language. An introduction to dynamic philology*. Cambridge: The MIT Press.

1.2 ZIPFŮV PROBLÉM PŘÍZVUKU

Hypotéza

„Nejvýraznějším rysem větného přízvuku je, že slova s nejvyšší frekvencí obvykle tento přízvuk nemají." (Zipf 1935/68: 131).

Ověřte tuto hypotézu.

Postup

Zipf se tento jev snažil demonstrovat na angličtině a němčině. Vy byste se proto mohli pokusit ověřit danou hypotézu na datech z jiných jazyků. Optimální jsou pro tento účel data ve formě mluveného jazyka, neboť psané texty je nutné nejprve nahlas přečíst a transkribovat. Běžný slovní přízvuk zde zanedbáváme, předmětem našeho zájmu je pouze hlavní přízvuk v rámci věty. Spočítejte všechna slova ve větách tvořících váš jazykový materiál a samostatně pak spočítejte přízvučná slova. Zpracujte si seznam veškeré slovní zásoby v textu, v němž bude uvedena celková frekvence jednotlivých slov a počet jejich výskytů v přízvučné formě. Pokud budou data vykazovat nějakou tendenci nebo jiný prvek pravidelnosti, vyjádřete takovou pravidelnost formálním způsobem, tj. jako funkci (aplikovanou na tato data). Nebude-li možné vysledovat žádnou zjevnou tendenci, pokuste se zjistit, za jakých podmínek by tato tendence mohla být patrná, tj. najděte takové charakteristiky textu, které umožní najít vzájemný vztah mezi frekvencí a preferencí k přízvučnosti.

Zipf sám korigoval své závěry pomocí výrazu „obvykle", tj. byl si vědom existence výjimek. Zaměřte se blíže na tyto výjimky a pokuste se je zdůvodnit z lingvistického hlediska.

Literatura

Zipf, G. K. (1935/1968). *The psycho-biology of language. An introduction to dynamic philology*. Cambridge: The MIT Press.

1.3 DISTINKTIVNOST PÍSMA

Problém

Vytvořte metodiku k měření rozlišovací schopnosti znaku.

Postup

Distinktivnost znaku nebo písmene je možné definovat pouze v rámci určitého systému. Např. distinktivnost písmového znaku ».« se tak liší podle toho, zda se považuje za jeden ze znaků na psacím stroji či v počítačovém fontu nebo za jeden ze tří znaků ».« »-« a » « Morseovy abecedy. Znaky či písmena sestávají z tahů (teček, rovných čar, křivek apod.). Nejprve si stanovte inventář těchto kresebných prvků v systému písma, které si přejete analyzovat. Následně určete vlastnosti tahů: délku (co nejvíce kategorií, které mají nějakou relevanci z hlediska rozlišovací funkce), pozici (vlevo-uprostřed-vpravo, nahoře-uprostřed-dole), sklon (vodorovný, svislý, šikmý – co nejvíce kategorií, které mají nějakou relevanci z hlediska rozlišovací funkce), otevřenost a šířku oblouků či půlkruhů v nejrůznějších relevantních směrech (např. sever, západ, jih, východ), tloušťku (je-li relevantní), nevybarvenost či vybarvenost (např. u čtverečků a kroužků) apod. Každé vlastnosti pak připište potřebné množství stupňů komplexity (v celočíselných hodnotách). Jednotlivé vlastnosti lze kombinovat, např. krátké horizontální rovné čáře vlevo nahoře bude přiřazena hodnota 1 apod. nebo budou hodnoty přiděleny způsobem rovná čára 1, nahoře 1, vlevo 1, krátká 1, vodorovná 1, přičemž tato čísla představují prvky vektoru. Sestavte vektor těchto charakteristik a všechny znaky mezi sebou porovnejte. Definujte distinktivnost znaku jako funkci jeho odlišností od všech ostatních znaků (např. jako průměr

všech rozdílů mezi položkami v rámci zkoumaného inventáře). Stanovte metodiku měření globální distinktivnosti systému písma jako celku.

K jiné variantě měření distinktivnosti lze dospět, budete-li mít k dispozici definici určitého fontu pomocí vektorů (např. definici fontu prostřednictvím Bézierových křivek). V takovém případě lze distinktivnost písmového znaku jednoduše určit na základě počtu a typu trajektorií, ze kterých se znak skládá, a počtu a (relativních) souřadnic jejich kontrolních bodů porovnáním příslušných popisů s popisem dalších znaků.

K podobnému výsledku lze dospět porovnáním pixelových (rastrových) definic jednotlivých znaků, je-li tímto způsobem definován daný font.

Porovnejte distinktivnost fontu latinky a řecké abecedy.

Zaměřte se na několik druhů písma, které se vyvinuly ze společného předchůdce, a věnujte pozornost zkoumání vývoje distinktivnosti z historického hlediska. Projděte jednotlivé fonty latinského písma, které váš textový procesor nabízí, a proveďte jejich vzájemné porovnání.

Porovnejte distinktivnost svého rukopisu s rukopisem svého kolegy.

Porovnejte distinktivnosti s dalšími vlastnostmi písma a rozhodněte, zdali mezi nimi existuje vzájemná závislost či alespoň souvztažnost.

Literatura

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

Antić, G., Altmann, G. (2005). On letter distinctivity. *Glottometrics 9,* 46–53.

1.4 ENTROPIE DISTINKTIVNOSTI SYSTÉMU PÍSMA

Problém

V návaznosti na část 1.3, „Distinktivnost písma", proveďte výpočet entropie této vlastnosti.

Postup

Rozhodněte se pro jednu z definic tahu v typech písma, o nichž bylo pojednáno v rámci předchozího problému. Nepřiřazujte tahům žádné hodnoty, ale zaznamenejte, kolikrát se v jednotlivých znacích vyskytne nějaký tah vyznačující se stejnými vlastnostmi. Získáte tak informaci o frekvencích (zastoupení) jednotlivých tahů. Na základě těchto frekvencí vypočtěte entropii distinktivnosti. Vysoká entropie je známkou vysoké distinktivnosti. Pokud písmo vykazuje velkou míru distinktivnosti, všechny typy tahů se v něm vyskytují se stejnou frekvencí.

Proveďte porovnání entropií distinktivnosti u různých fontů typu „Roman". Tyto výsledky porovnejte s Morseovou abecedou a Braillovým písmem z hlediska jejich relativních entropií.

Distinktivnost může na druhou stranu snížit existence znaků sestávajících z dlouhé řady shodných tahů. V Morseově abecedě a tzv. ogamovém písmu se vyskytuje až pět shodných tahů v řadě za sebou. Vyjádřete entropii distinktivnosti u těchto dvou písmových systémů.

V případě potřeby můžete také uplatnit Shannonovu definici entropie.

Literatura

žádná

1.5 KOMPLEXITA PÍSMA 2

Problém

V knize *Kvantitativní lingvistika. Vybrané problémy 1* (Strauss et al. 2014, kap. „Komplexita písma") jsme komplexitu pojímali jako čistě grafickou vlastnost znaku, kterou lze aplikovat na jakýkoli druh písma. Jelikož však každý pojem lze operacionalizovat různým způsobem – definice nemají pravdivostní hodnotu – představíme si zde ještě jiné možnosti. Lze je uplatnit pouze u alfabetických druhů písma, neboť se týkají vztahu mezi písmeny a grafémy na straně

jedné a fonémy na straně druhé. Aktuální úkol spočívá ve zpracování většího množství jazyků.

Postup

Od *fonému* ke *grafému*: Pracujte s kompletním inventářem fonémů určitého jazyka. Ke každému fonému uveďte všechna písmena nebo skupiny písmen, které jej mohou reprezentovat. Například anglické /m/ mohou reprezentovat písmena nebo skupiny písmen <*gm, m, mb, me, mm, mme, mn, mp, nm*>. Velikost této ortografické množiny je náznakem ortografické neurčitosti fonému. Proveďte úplné zpracování jednoho jazyka a vyjádřete tento typ komplexity použitím informačněteoretických způsobů měření neurčitosti (srov. Altmann, Fan 2008), případně vypracujte nové postupy takového měření. V přehledu referenční literatury níže najdete některé práce, které mohou posloužit jako příklady tohoto druhu analýzy.

Od písmene ke grafému: Bosch et al. (1974: 178) definují grafém jako „písmeno nebo uskupení písmen, jež je ve fonologické transkripci realizováno jako jeden samostatný foném". Písmena se mohou vyskytovat ve formě různých grafémů. Vypočítejte distribuci písmen, které se mohou vyskytovat v 1, 2, 3, … grafémech. Prokažte, že se jedná o monotónně klesající distribuci, a pokuste se určit její formu. Pokud nevykazuje monotónní charakteristiky, pokuste se zjistit, proč tomu tak je, a své výsledky zobecnit. Altmann a Fan (2008) používají termín „grafemická zátěž písmen", tato vlastnost současně vypovídá o formální komplexitě psaní. Metodu, kterou zde prezentujeme, lze také uplatnit na ideografická písma, v nichž roli písmen sehrávají různé druhy tahů. Příslušná forma jejich distribuce nám umožňuje definovat indikátory komplexity. Za použití výsledků níže uvedených studií proveďte analýzu některých dalších jazyků.

Literatura

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

Berndt, R. S., Reggia, J. A., Mitchum, C. C. (1987). Empirically derived probabilities for grapheme-to-phoneme correspondences in English. *Behaviour Research Methods, Instruments & Computers 19,* 1–9.

Best, K.-H., Altmann, G. (2005). Some properties of graphemic systems. *Glottometrics 9,* 29–39.

Bosch, A. v. d., Content, A., Daelemans, W., Gelder, B. de (1974). Measuring the complexity of writing systems. *Journal of Quantitative Linguistics 1*(3), *178–188*.

Fry, E. (2004). Phonics: a large phoneme-grapheme frequency count revised. *Journal of Literacy Research 36(1),* 85–98.

Grzybek, P., Kelih, E. (2003). Graphemhäufigkeiten (am Beispiel des Russischen). Teil I: Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen. *Anzeiger für Slavische Philologie 31,* 131–162.

Hanna, P. R., Hanna, J. S., Hodges, R. E., Rudorf, E. H. (1966). *Phoneme-grapheme correspondences as cues to spelling improvement*. Washington, D.C.: U.S. Department of Health, Education, and Welfare.

Patterson, K. E., Morton, J. (1985). From orthography to phonology: An attempt at an old interpretation. In: Patterson, K. E., Marshall, J., Coltheart, W. (eds.), *Surface dyslexia: neuropsychological and cognitive studies of phonological reading*. London: Erlbaum.

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

1.6 KANONICKÉ ŘEČOVÉ SEGMENTY

Problém

Určete rankovou a spektrální distribuci kanonicky transkribovaných řečových segmentů.

Postup

Zvolte si text obsahující minimálně 1 000 slov (tokenů). Pokud provádíte analýzu psaného jazyka, jeho ortografický zápis musí být nejprve (za využití počítačového programu nebo ručně) transkribován prostřednictvím fonetických nebo alofonických symbolů. Následně si zvolte minimálně dvě klasifikační třídy hlásek (např. V = vokál a K = konsonant). Tuto klasifikaci lze rozšířit, např. na vokály, konsonantny, klouzavé hlásky („glidy"), polovokály, redukované vokály atd. Přiřaďte symboly označující hlásky v textu k těmto třídám (tj. označte je odpovídajícími symboly). Nyní určete počty tokenů jednotlivých typů (tříd) v textu a výsledné počty uspořádejte podle rankové distribuce a odpovídajícího spektra. Pokuste se zjistit, která teoretická rozdělení by bylo možné úspěšně aplikovat na vaše data.

Jako dobré východisko mohou posloužit data Galea a Sampsona (1995), kteří zpracovali přepis anglických hlásek na základě tří kategorií, V – vokál, R – redukovaný vokál, K – konsonant, čímž získali typy jako např. VKV, VKKR-KRKV, VRRKKV apod. Data extrahovaná z jednotlivých textů uvádí tabulka 1.6.1, kde x = počet výskytů a n_x = počet typů vyskytujících se přesně x krát. Následně převeďte spektrální distribuci na rankovou distribuci a určete příslušnou teoretickou distribuci.

Vypočtěte různé indikátory, které popisují tato i vaše vlastní data (srov. Popescu et al. 2009), a pokuste se nalézt nějaký rys, který by byl společný všem těmto datům. Opatřete si pokud možno data z různých jazyků a tyto jazyky mezi sebou porovnejte.

TABULKA 1.6.1, Spektrální distribuce kanonických řečových segmentů v angličtině (Gale, Sampson 1995)

x	n_x	x	n_x	x	n_x	x	n_x
1	120	20	3	46	1	257	1
2	40	21	2	47	1	339	1
3	24	23	3	50	1	421	1

x	n_x	x	n_x	x	n_x	x	n_x
4	13	24	3	71	1	456	1
5	15	25	3	84	1	481	1
6	5	26	2	101	1	483	1
7	11	27	2	105	1	1140	1
8	2	28	1	121	1	1256	1
9	2	31	2	124	1	1322	1
10	1	32	2	146	1	1530	1
12	3	33	1	162	1	2131	1
14	2	34	2	193	1	2395	1
15	1	36	2	199	1	6925	1
16	1	41	3	224	1	7846	1
17	3	43	1	226	1		
19	1	45	3	254	1		

Literatura

Gale, W. A., Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics 2(3)*, 217–237.

Popescu, I.-I. et al. (2009). *Aspects of word frequencies*. Berlin, New York: de Gruyter.

1.7 FONETICKÉ SROVNÁNÍ PŘÍBUZNÝCH JAZYKŮ

Problém

Prostřednictvím kvantitativních konceptů vyjádřete fonetickou podobnost příbuzných jazyků, případně nějakého staršího a moderního jazyka, který se z něj vyvinul, anebo určete míru asimilace výpůjček.

Postup

Při porovnávání dvou příbuzných jazyků z hlediska jejich fonetické podobnosti (či odlišnosti) pracujte s obecným fonetickým systémem a hodnotami rozdílů mezi jednotlivými hláskami vyjádřenými alespoň na ordinální škále. Takové měření rozdílů může být například založeno na počtu odlišností v místě a způsobu artikulace nebo na množství různých distinktivních rysů dvou hlásek. Následně si vytvořte vzorek náhodně vybraných příbuzných slov nebo použijte příbuzná slova ze Swadeshova (1964) seznamu základní slovní zásoby. Porovnejte rozdíly mezi jednotlivými odpovídajícími si hláskami v každém slově – provádějte rovněž kvantitativní hodnocení ztráty či přidání hlásek, epentezí apod. – a vyjádřete fonetické rozdíly mezi dvěma příbuznými jazyky jako průměr všech odlišností.

(1) Porovnejte tímto způsobem dva románské jazyky, latinu a jazyky, které se z ní vyvinuly, některé slovanské jazyky apod.

(2) Zjistěte, zda prostorová vzdálenost mezi příbuznými jazyky koreluje s fonetickými rozdílnostmi.

(3) Proveďte kvantitativní vyhodnocení fonetické změny výpůjček nebo jednotlivých hlásek ve vypůjčených slovech v cílovém jazyce.

(4) Nepoužívejte měřítka, která berou v úvahu pouze přítomnost změny, ale nezohledňují její fonetickou povahu, např. Levenshteinovu vzdálenost.

Literatura

a Campo, F., Geršić, S., Naumann, C. L., Altmann, G. (1985). Subjektive Lautähnlichkeit. *Beiträge zur Phonetik und Linguistik 50,* 101–120.

Augst, G. (1971). Über die Kombination von Phonemsequenzen bei Monemen. *Linguistische Berichte 11,* 37–47.

Austin, W. M. (1957). Criteria for phonetic similarity. *Language 33,* 538–544.

Batóg, T., Steffen-Batogowa, M. (1980). A distance function in phonetics. *Lingua Posnaniensis 23,* 47–58.

Geršić, S. (1971). *Mathematisch-statistische Untersuchungen zur phonetischen Variabilität, am Beispiel von Mundartaufnahmen aus der Batschka*. Göppingen: Kümmerle.

Grimes, J. E., Agard, F. B. (1959). Linguistic divergence in Romance. *Language 35*, 598–604.

Grotjahn, R. (1980). Zur Quantifizierung der Schwierigkeit des Sprechbewegungsablaufs. In: Grotjahn, R., Hopkins, E. (Hrsg.), *Empirical research on language teaching and language acquisition.* Bochum: Brockmeyer, 199–231.

Ladefoged, P. (1970). The measurement of phonetic similarity. *Statistical Methods in Linguistics* 6, 23–32.

Lehfeldt, W. (1978). Zur Messung der phonetischen Lautdifferenz. Eine begriffskritische Untersuchung. *Glottometrika 1,* 26–45.

Lehfeldt, W. (1980). Zur numerischen Erfassung der Schwierigkeit des Sprechbewegungsablaufs. *Glottometrika 2,* 44–61.

Levenštejn, V. I. (1965). Dvoičnyje kody s ispravleniem vypadenij, vstavok i zameščenij simvolov. *Doklady Akademii Nauk SSSR 163(4)*, 845–848. [Vyšlo v angličtině jako: Levenshtein, V. I., Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10 (1966),* 707–710.]

Lindner, G. (1980). Lautfolgestrukturen im Deutschen. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 33,* 468–477.

Peterson, G. H., Harary, F. (1961). Foundations of phonemic theory. In: Jakobson, R. (Hrsg.), *Structure of language and its mathematical aspects*. Providence, Rhode Island: American Mathematical Society, 139–165.

Swadesh, M. (1964). Linguistics as an instrument of prehistory. In: Hymes, D. (ed.), *Language in Culture and Society: A Reader in Linguistics and Anthropology*. New York: Harper and Row, 575–584.

Tolstaja, S. M. (1983). Fonologičeskoe rasstojanie i sočetaemosť soglasnych v slavjanskich jazykach. *Voprosy jazykoznanija 3,* 66–81.

1.8 FONETICKÁ STRUKTURA SLOVA

Problém

(a) Vyhodnoťte průměrný fonetický rozdíl mezi následnými hláskami v první stovce nejfrekventovanějších slov určitého jazyka.

(b) Vyhodnoťte průběh fonetických rozdílů v každém slově (jako rozdíl sousedních hlásek) a vytvořte na tomto základě klasifikaci slov.

Postup

Použijte způsob měření fonetických rozdílů popsaný v doporučené literatuře uvedené v části 1.7, „Fonetické srovnání příbuzných jazyků", nebo vypracujte svou vlastní metodiku měření. Sestavte si velký výběrový soubor slov z frekvenčního slovníku (nejfrekventovanější slova) a ověřte, zda existuje nějaká souvislost mezi průměrným fonetickým rozdílem v daném slově a frekvencí tohoto slova. Pokud byly Zipfovy (1949) domněnky správné, pak by mezi těmito kvantitami měla existovat určitá rovnováha nebo závislost. Mluvčí usiluje o malý průměrný rozdíl (ekonomika produkce), zatímco posluchač o velký (ekonomika dekódování). Ověřte, zda by mohl být tento vztah vyjádřen funkcí gama. Pokud ano, formulujte teoretické zdůvodnění pro funkci gama.

K vyřešení problému (b) je třeba provést analýzu co největšího počtu slov ze slovníku a pozorovat průběh odlišností. Slova nejprve rozdělte podle jejich délky (z hlediska hlásek) a v rámci kategorií délky pak podle průběhu vzorů rozdílnosti.

Zaměřte se na počet vzorů a uveďte, zda existuje pravidelné rozdělení frekvence (i) v rámci konkrétní délkové kategorie a (ii) v jazyce jako takovém. Pokud ano, stanovte teoretické rozdělení a zdůvodněte jej.

Literatura

Srov. problém 1.7

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge: Addison-Wesley.

1.9 FONETICKÁ DEFORMACE VÝPŮJČEK

Problém

Vypůjčená slova jsou obvykle foneticky modifikována. Vyjádřete rozsah fonetické deformace.

Postup

Nejprve si zmapujte různé způsoby práce s fonetickou podobností. Tímto tématem se zabývá nepřeberné množství odborné literatury (pro zajímavost si do internetového vyhledávače zadejte klíčové slovo „phonetic similarity"). Nejčastěji je na tento pojem možné narazit v dialektologii (srov. např. Goebl 1984). Deformací se nemíní pouze nahrazení, ale také eliminace a přidání dalších hlásek. Nejoblíbenějším měřítkem různosti je proto Levenshteinova vzdálenost, byť mnohé další typy měření vzdálenosti mohou fungovat přinejmenším stejně.

Připravte si seznam foneticky přepsaných vypůjčených slov ve zdrojovém i cílovém jazyce. Následně si sestavte tabulku obsahující škálované hodnoty každé vlastnosti pro všechny hlásky (v obou jazycích). Příklad takového škálování u předozadní dimenze: 1. labiála, 2. labiodentála, 3. alveolára, 4. palatála, 5. uvulára, 6. laryngála. Nakonec porovnejte každé slovo ze zdrojového jazyka

s příslušným slovem v cílovém jazyce hlásku po hlásce. Součet deformací ve slově představuje hodnotu proměnné *D*.

(1) Stanovte frekvenční distribuci proměnné *D*. Pokud máte vypočteny fonetické rozdíly na poměrové škále, vytvořte intervaly pro *D* a použijte spojité rozdělení. Dochází k monotónnímu poklesu distribuce? Pokud ano, proč? Pokud ne, proč ne?

(2) Zaměřte se na jednu konkrétní hlásku zdrojového jazyka. Ne vždy se mění na tutéž cílovou hlásku, její deformace může mít různou podobu, např. hláska /a/ ve zdrojovém jazyce se může změnit na /a/, /ą/, /o/ a /#/ v cílovém jazyce. V takovém případě má její diverzifikace hodnotu 4. U každé zdrojové hlásky zjistěte počet cílových hlásek, v něž se může změnit (včetně eliminace). Stanovte distribuci počtu diverzifikací hlásek. Parametry této distribuce použijte k změření fonetické vzdálenosti mezi zdrojovým a cílovým jazykem.

Literatura

a Campo, F., Geršić, S., Naumann, C. L., Altmann, G. (1989). Subjektive Ähnlichkeit deutscher Laute. *Glottometrika 10,* 46–70.

Afendras, E. A., Tzannes, N. S., Trépanier, J. G. (1973). Distance, variation and change in phonology: stochastic aspects. *Folia Linguistica 6,* 1–27.

Austin, W. M. (1957). Criteria for phonetic similarity. *Language 33,* 538–544.

Batóg, T., Steffen-Batogowa, M. (1960). A distance function in phonetics. *Lingua Posnaniensis 23,* 47–58.

Benzecri, J. P. (1970). Sur l'analyse des matrices de confusion. *Revue de statistique appliquée 18,* 5–63.

Bruce, D., Murdock, B. B. Jr. (1968). Acoustic similarity effects on memory for paired associates. *Journal of Verbal Learning and Verbal Behaviour 7,* 627–631.

Cucchiarini, C. (1993). *Phonetic transcription: a methodological and empirical study.* Nijmegen, Diss.

Geršić, S. (1971). *Mathematisch-statistische Untersuchungen zur phonetischen Variabilität, am Beispiel von Mundartaufnahmen aus der Batschka*. Göppingen: Kümmerle.

Geršić, S., Naumann, C. L., Altmann, G. (1985). Subjektive Lautähnlichkeit. *Beiträge zur Phonetik und Linguistik 50,* 101–120.

Goebl, H. (1984). *Dialektometrische Studien*. 3. vols. Tübingen: Niemeyer.

Grimes, J. E., Agard, F. B. (1959). Linguistic divergence in Romance. *Language 35,* 598–604.

Grotjahn, R. (1980). Zur Quantifizierung der Schwierigkeit des Sprechbewegungsablaufs. In: Grotjahn, R., Hopkins, E. (eds.), *Empirical research on language teaching and language acquisition*. Bochum: Brockmeyer, 199–231.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Groningen, Diss.

Klatt, D. H. (1968). Structure of confusions in short-term memory between English consonants. *The Journal of the Acoustic Society of America 44,* 401–407.

Kondrak, G. (2003). Phonetic alignment and similarity. *Computers and the Humanities 37(3),* 273–291.

Lehfeldt, W. (1980). Zur numerischen Erfassung der Schwierigkeit des Sprechbewegungsablaufs. *Glottometrika 2,* 44–61.

Lindner, G. (1975). *Der Sprechbewegungsablauf. Eine phonetische Studie des Deutschen*. Berlin: Akademie-Verlag.

Lindner, G. (1980). Lautfolgestrukturen im Deutschen. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 33,* 468–477.

Łobacz, P. (1981). Classification of Polish consonantal fonéms on the basis of a subjective similarity test. *Speech Analysis and Synthesis 5,* 97–120.

Łobacz, P. (1981). Distances between Polish consonantal fonéms in tests with real and nonsense words. *Speech Analysis and Synthesis 5,* 121–138.

Miller, G. A., Nicely, P. E. (1955). An analysis of perceptual confusion among consonants. *Journal of the Acoustical Society of America 27,* 338–352.

Mohr, B., Wang, W. S.-Y. (1968). Perceptual distance and the specification of phonological features. *Phonetica 18*, 31–45.

Nakatani, L. H. (1972). Confusion-choice model for multidimensional psychophysics. *Journal of Mathematical Psychology 9,* 104–127.

Nerbonne, J., Heeringa, W., Kleiweg, P. (1999). Edit distance and dialect proximity. In: Sankoff, D., Kruskal, J. (eds.), *Time warps, string edits and macromolecules: the theory and practice of sequence comparison: V–XV*. Stanford, CA: CSLI (2nd ed.).

Nerbonne, J., Kleiweg, P. (2007). Toward a dialectological yardstick. *Journal of Quantitative Linguistics 14*(2–3), 148–166.

Peterson, G. H., Harary F. (1961). Foundations of phonemic theory. In: Jakobson, R. (ed.), *Structure of language and its mathematical aspects*. Providence, Rhode Island, 139–145.

Singh, S. (1966). Crosslanguage study of perceptual confusion of plosive fonéms in two conditions of distortion. *Journal of the Acoustical Society of America 40,* 635–656.

Singh, S. (1971). Perceptual similarities and minimal phonemic difference. *Journal of Speech and Hearing Research 14*, 113–124.

Singh, S., Black J. W. (1966). Study of twenty-six intervocalic consonants as spoken and recognized by four language groups. *Journal of the Acoustical Society of America 39,* 372–387.

Singh, S., Frank, D. C. (1972). A distinctive feature analysis of the consonantal substitution pattern. *Language and Speech 15,* 209–218.

Thürmann, E. (1974). Phonetische Ähnlichkeit, distinktive Merkmale und auditive Dimensionen. Ein Bericht. *Hamburger Phonetische Beiträge 13,* 163–192.

Tolstaja, S. M. (1968). Fonologičeskoe rasstojanie i sočetaemosť soglasnych v slavjanskich jazykach. *Voprosy jazykoznanija,* 66–81.

Wilson, K. V. (1963). Multidimensional analyses of confusions of English consonants. *The American Journal of Psychology 76,* 89–95.

Yokoyama, S., Itahashi, S. (1979). On the distance of Japanese words based on distinctive features and a second-order model. *Glottometrika 2,* 62–81.

2 Gramatika

2.1 FENKOVA HYPOTÉZA

Hypotéza

Jednotky s vysokou frekvencí se ve větě nacházejí v její přední části, zatímco méně časté jednotky jsou spíše vzadu. Důvodem je, že „… multifunkční slova mají tendenci hromadit se v první části věty…“ (Fenk-Oczlon, Fenk 2002). Ověřte tuto hypotézu.

Postup

(1) Zpracujte si frekvenční seznam slov lemmatizovaného textu. Místo počítání frekvence lze použít frekvenční slovník.

(2) Rozdělte daný text na dílčí celky obsahující věty o stejné délce, tj. skládající se ze 2, 3, 4,… slov.

(3) V každém celku nahraďte příslušná slova jejich frekvencemi (použijte buď frekvenci z daného textu, nebo příslušnou hodnotu z frekvenčního slovníku). Neopomíjejte interpunkční znaménka, mohou sehrávat podstatnou roli.

(4) Pro každou pozici v jednotlivých celcích s různě dlouhými větami vypočtěte pro každou pozici zvlášť její průměrnou frekvenci.

(5) Formulujte hypotézu o sekvenci průměrných frekvencí v průběhu věty. Podle Fenkovy hypotézy by měla frekvence monotónně klesat. Pokud vaše data tuto hypotézu potvrdí, pokuste se ji dále rozpracovat. Bude-li to možné, navrhněte funkci vyjadřující pokles příslušných hodnot. Jak se dané parametry mění s délkou věty?

Pokud se platnost hypotézy nepodaří ověřit, navyšte velikost vašeho datového souboru nebo formulaci hypotézy upravte. Zohledněte postavení interpunkčních znamének.

Tato hypotéza již byla ověřována na angličtině, a proto by bylo vhodné zaměřit se zde spíše na nějaký jiný než indoevropský jazyk. Budete-li moci porovnávat několik jazyků, snažte se do dané hypotézy zakomponovat hraniční podmínky, tj. určit faktory, které jsou zodpovědné za pozorované rozdíly mezi výsledky u dat z různých jazyků.

Dospějete-li ke kladným výsledkům, využijte příslušné parametry funkce k typologickým účelům. Mohou být jazyky charakterizovány těmito parametry?

Je u některých jazyků možné vysledovat specifický průběh frekvenčních sekvencí? Pokud ano, mají tyto jazyky/průběhy nějaké společné rysy? Pokud vypozorujete nějakou pravidelnost, pokuste se ji popsat nějakou matematickou funkcí a teoreticky ji odvodit na základě určitých syntaktických vlastností daného jazyka. Jak se tato funkce mění s rostoucí délkou vět? Proveďte analýzu dat zvlášť pro jednotlivé skupiny vět podle délky.

Literatura

Fenk-Oczlon, G., Fenk, A. (2002). Zipf´s tool analogy and word order. *Glottometrics 5,* 22–28.

Fenk, A., Fenk-Oczlon, G. (2002a). 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.

2.2 ZIPFOVA HYPOTÉZA TÝKAJÍCÍ SE ADVERBIÍ (1)

Hypotéza

„… adverbia času jsou v průměru méně nezávislá, a proto kratší než adverbia místa." (Zipf 1935/68: 242). Ověřte tuto hypotézu.

Postup

Zpracujte si seznam všech adverbií místa a času, která naleznete v učebnicích nebo gramatikách více různých jazyků. Vypočtěte průměrné délky těchto dvou skupin a porovnejte je např. pomocí t-testu. Můžete potvrdit Zipfovu hypotézu? Pokud ne, můžete stanovit nějaké hraniční podmínky, na základě jejichž splnění by bylo možné přijmout tuto jednoduchou hypotézu? Najdete nějaké jazyky, v nichž by se dvě takové skupiny vyznačovaly přibližně stejnou průměrnou délkou? Pokud ano, určete konkrétní vlastnosti takových jazyků. Je nutné modifikovat danou hypotézu? Ke každému adverbiu přiřaďte jeho frekvenci výskytu vypočtenou z korpusu nebo udanou frekvenčním slovníkem. Je mezi nimi možné vysledovat nějaké souvislosti?

Věnujte pozornost jiným jazykům než těm, které jsou nejčastěji předmětem zkoumání, jako např. angličtina. Zpřesněte formulaci hypotézy tím, že změříte rozsah kotextové závislosti adverbií. Vzniká tím samostatný, netriviální problém, který bude pravděpodobně možné vyřešit pouze ve spojení se slovesnou valencí.

Literatura

Zipf, G. K. (1935/1968). *The psycho-biology of language. An introduction to dynamic philology*. Cambridge: The MIT Press.

2.3 ZIPFOVA HYPOTÉZA TÝKAJÍCÍ SE ADVERBIÍ (2)

Hypotéza

„… adverbia času jsou v průměru méně nezávislá, a proto kratší než adverbia místa." (Zipf 1935/68: 242). Ověřte tuto hypotézu.

Postup

Zipfova hypotéza vychází z předpokladu, že adverbia času jsou méně nezávislá (např. na slovesech) než adverbia místa. Tento předpoklad je třeba ověřit jako takový. K měření nezávislosti lze využít polytextualitu.

Můžete opětovně použít materiál z části 2.2, „Zipfova hypotéza týkající se adverbií (1)". Jednotlivé kontexty s výskytem adverbií roztřiďte podle nějakého sémantického anebo pragmatického kritéria. Počet takto vytvořených tříd by měl být vyšší než čtyři. Následně každou třídu, v níž se daná adverbia vyskytují, označte číslem a toto číslo přidělte jednotlivým adverbiím (polytextualita). Obě skupiny porovnejte prostřednictvím chí-kvadrát testu nebo t-testu.

Literatura

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

Köhler, R. (2006). Frequenz, Kontextualität und Länge von Wörtern. Eine Erweiterung des synergetisch-linguistischen Modells. In: Rapp, R., Sedlmeier, P., Zunker-Rapp, G. (eds.), *Perspectives on Cognition*. Lengerich: Pabst Science Publishers, 327–338.

Zipf, G. K. (1935/1968). *The psycho-biology of language. An introduction to dynamic philology*. Cambridge: The MIT Press.

2.4 POMOCNÁ SLOVA (AUXILIÁRY)

Problém

Je známo, že pomocná slova (auxiliáry) jsou v řeči i textu nejfrekventovanější. K této skutečnosti se váže řada různých hypotéz, např. že „… existuje výrazná korelace mezi vysokou frekvencí a pomocnou funkcí slova." (Krug 2001: 312). Vědci však v této souvislosti uvažují většinou indoevropské jazyky. Zaměřte se blíže na následující úkoly:

Postup

Nejprve si definujte třídu pomocných slov (auxiliár) v jazyce, který si zvolíte. Definici formulujte co nejpřesněji a vytvořte metodu pro škálování míry pomocné funkce slova.

(1) V několika textech porovnejte rankovou distribuci auxiliár v nějakém silně syntetickém a oproti tomu silně analytickém jazyce, např. ve slovanském, respektive polynéském jazyce. Demonstrujte existenci diametrálních rozdílů mezi ranky (= skóry na vaší škále nezávislosti) u těchto pomocných slov. Tento rozdíl prokažte provedením statistického testu.

(2) Navrhněte index vyjadřující rozsah využívání auxiliár v daném jazyce. Dbejte na jeho jednoduchost a porovnatelnost. Charakterizujte zkoumané jazyky. Jednotlivé ranky relativizujte (= vydělte každý rank velikostí inventáře V), aby bylo možné porovnávat texty různé délky. Proveďte neparametrický test různosti daných jazyků.

Literatura

Krug, M. G. (2001). Frequency, iconicity, categorization: evidence from emerging modals. In: Bybee, J., Hopper, P., *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 309–335.

2.5 VALENCE A FREKVENCE V TEXTU

Problém

Jsou slovesa s valencí x (x = 1, 2, 3, … , n) rozdělena v rámci určitého korpusu pravidelně?

Postup

Zjistěte frekvenci všech sloves ve zvoleném korpusu. Poté stanovte valenci každého slovesa v rámci daného korpusu, tj. nepoužívejte valenční slovník. Lze pozorovat, že slovesa s vysokou valencí vykazují vyšší frekvenci výskytu než nízkovalenční slovesa. V lexiku se na druhou stranu nachází více sloves s nízkou frekvencí než s vysokou valencí. Při absenci jakéhokoli trendu bychom tedy mohli očekávat rovnoměrné rozdělení. Uveďte, zda se pozorovaná empirická distribuce $\{P_x\}$ – kde x = valence, $f(x)$ = frekvence všech sloves s valencí x – liší od rovnoměrného rozdělení, a pokud ano, jaký typ rozdělení představuje vhodný model pozorované struktury? Nejprve určete distribuci empiricky (pomocí statistického softwaru, který nabízí velký počet teoretických rozdělení pravděpodobnosti), a poté svůj přístup zdůvodněte, tj. odvoďte jej z lingvistických argumentů.

Jako sekundární výstup své analýzy pak určete, zda postavení sloves ve větách nějak souvisí s mírou valence.

Literatura

žádná

2.6 VALENCE A RANKOVÉ POŘADÍ

Problém

Ranková distribuce valence se řídí obecným rankovým rozdělením či funkcí. Ověřte tuto hypotézu.

Postup

Převeďte si empirickou distribuci v části 2.5, „Valence a frekvence v textu", na distribuci rankového pořadí, tj. neuvažujte valenci, ale přiřaďte rank 1 slovesu s nejvyšší frekvencí, rank 2 druhému nejfrekventovanějšímu slovesu atd. Získáte tak

rankovou distribuci sloves (nikoli valenci!). Pokuste se zjistit příslušnou formu distribuce empiricky i teoreticky.

Následně z distribuce extrahujte slovesa s valencí 1, seřaďte je sestupně podle frekvence a určete distribuci nebo prostou funkci pro tuto řadu. Poté vyčleňte slovesa s valencí 2 a proveďte totéž. Pokračujte až po nejvyšší valenční třídu s minimálně pěti slovesy. Poté, co identifikujete všechny distribuce nebo funkce (sekvence), porovnejte je a formulujte novou hypotézu o formě distribuce ve vztahu k její závislosti na valenční třídě. Jedná se vždy o tutéž funkci s různými parametry nebo potřebujete různé funkce pro různé valenční třídy?

Lze tento postup použít jako kritérium obhajoby tradičního připisování valenční hodnoty slovesům?

Literatura

žádná

2.7 DIVERZIFIKACE PÁDŮ V UGROFINSKÝCH JAZYCÍCH

Problém

Vypočtěte diverzifikační konstantu (viz část 6.3, „Diverzifikační konstanta“) pro pád substantiv v některých ugrofinských jazycích.

Postup

Zvolte si minimálně tři ugrofinské jazyky a 10 textů v každém z těchto jazyků. Stanovte co nejpřesněji, které sufixy vyjadřují u substantiv pádové vztahy. Sdružte všechny alomorfy, které jsou výsledkem vokálové harmonie, a neberte v potaz polysémii nebo polyfunkčnost jednotlivých afixů. Na základě absolutních frekvencí sestavte rankovou posloupnost afixů a demonstrujte, že (1) tuto rankovou sekvenci je možné modelovat prostřednictvím Popescovy funkce

$f(r) = 1+a^*\exp(-r/b)$, kde r je rank a $f(r)$ je absolutní frekvence ranku r, a (2) vypočtěte diverzifikační konstantu c

$$c = \frac{R + f_{max} - f_{min} + 1 - L}{h}$$

pro každý jazyk zvlášť a demonstrujte jejich značnou podobnost. Pomocí tabulky 6.3.1 a vzorců uvedených v části 6.3, „Diverzifikační konstanta", určete, do které kategorie jevů tento případ spadá.

Literatura

Best, K.-H. (2009). Diversifikation des Phonems /r/ im Deutschen. *Glottometrics 18*, 26–31.

Laufer, J., Nemcová, E. (2009). Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18,* 13–35.

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18,* 32–39.

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17,* 97–111.

Roos, U. (1991). Zur Diversifikation der japanischen Postposition *ni*. In: Rothe, U. (ed.) (1991), *Diversification processes in language: grammar.* Hagen: Rottmann, 75–82.

Rothe, U. (ed.) (1991). *Diversification process in language: grammar.* Hagen: Rottmann.

Sanada, H. (2009). Diversification of postpositions in Japanese. *Glottometrics* 19, 70–79.

2.8 VALENCE A KOMPOZITA

Hypotéza

Čím má sloveso větší valenci, tím více kompozit vytváří.

Postup

Uvažujte valenci jako počet komplementů (případně počet všech doplnění, tj. komplementů a adjunktů), které sloveso vyžaduje k vytvoření úplné věty. Například věta *Já jsem včera viděl v kině dobrý film* má dva komplementy (*já* a *dobrý film*) a dva adjunkty (*včera* a *v kině*). Z valenčního slovníku jazyka, který jste si vybrali pro analýzu, náhodně vyberte minimálně 300 sloves. Každé sloveso označte hodnotou vyjadřující jeho valenci. V běžném výkladovém slovníku (nebo v nějakém internetovém zdroji) zjistěte u každého slovesa počet kompozit, které vytváří. Vypočtěte průměrný počet kompozit u každé valenční třídy. Následně se explorativním způsobem pokuste určit funkci vyjadřující závislost kompozicionality na valenci: KOMP = *f*(VAL). Pokud se vám to podaří, danou funkci zdůvodněte. Sestavte diferenciální rovnici, která vede k dané funkci, a zdůvodněte ji. Svůj výsledek znázorněte ve formě diagramu. Zakomponujte jej do rámce synergetické lingvistiky (Köhler 1986, 2002, 2005) a demonstrujte, že danou funkci lze odvodit ze sjednocené teorie (Wimmer, Altmann 2005).

Literatura

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

Köhler, R. (ed.) (2002). *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik*. [Dostupné z: http://ubt.opus.hbz-nrw.de/volltexte/2004/279 (cit. 12. prosince 2008).]

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, 760–774.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, 791–807.

2.9 VALENCE A ODVOZOVÁNÍ

Hypotéza

Čím má sloveso větší valenci, tím je produktivnější, tj. tím více odvozenin může vytvářet. Ověřte tuto hypotézu.

Postup

Za použití valenčního slovníku nebo odpovídajících elektronických zdrojů na internetu si sestavte náhodný výběrový soubor čítající 200 jednoduchých, tj. neodvozených sloves a určete počet jejich valencí. Vytvořte tabulku, kde v prvním sloupci budou jednotlivá slovesa a ve druhém počet jejich valencí. Zjistěte všechny odvozeniny těchto sloves (zohledněte rovněž odvozování prostřednictvím konkatenace – např. německé *Ver/**beug**/ung*, maďarské *meg/**lep**/etés*, slovenské *roz/**del**/enie* – a alternace hlásek (např. ablaut nebo umlaut), nikoli však konverzi nebo odvozování nulovým morfémem). Počet odvozenin uveďte ve třetím sloupci tabulky. Poté nejprve vypočtěte korelaci mezi valenčními a derivačními hodnotami. Používáte-li elektronický tabulkový procesor, data lze snadno graficky znázornit, a získat tak představu o podobě závislostní funkce. Pokud zaznamenáte příliš velký rozptyl, doplňte výběrový soubor o dalších 100 slov. Dospějete-li k zajímavému výsledku, najděte funkci vyjadřující závislost. V ideálním případě odvoďte tuto funkci na základě teoretické úvahy, jinak se pokuste postupovat induktivně.

Pokud tento vzájemný vztah existuje, určitě bude záviset na míře odvozování v daném jazyce. Tuto vlastnost zaveďte jako nezávislou proměnnou, aby bylo možné provádět přesnější predikce. Koncept „odvozovatelnosti" vyjádřete pomocí varianty Greenberg-Krupových indexů (viz referenční literatura níže).

Literatura

Allerton, D. (1982). *Valency and the English verb*. London, New York: Academic Press.

Emons, R. (1978). *Valenzgrammatik für das Englische. Eine Einführung*. Tübingen: Niemeyer.

Engel, U. et al. (1983). *Valenzlexikon deutsch-rumänisch*. Heidelberg.

Greenberg, J. H. (1960/1990). A quantitative approach to the morphological typology of languages. In: Denning, K., Kemmer, S. (eds.), *On Language: Selected Writings of Joseph H. Greenberg.* Stanford, California: Stanford University Press, 3–25.

Hajič, J. (1998). Building a syntactically annotated corpus: The Prague Dependency Treebank. In: Hajičová, E. (ed.), *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*. Praha: Karolinum, 106–132.

Helbig, G. (Hrsg.) (1971). *Beiträge zur Valenztheorie*. The Hague, Paris: de Gruyter.

Helbig, G. (1992). *Probleme der Valenz- und Kasustheorie*. Tübingen: Niemeyer.

Helbig, G., Schenkel, W. (1969/83). *Wörterbuch zur Valenz und Distribution deutscher Verben.* VEB Bibliographisches Institut.

Hudson, R. (1993). Recent developments in dependency theory. In: Jacobs, J., Stechow, A. v., Sternefeld, W., Vennemann, T. (eds.), *Syntax. Ein internationales Handbuch zeitgenössischer Forschung*. Berlin, New York: de Gruyter, 329–338.

Krupa, V. (1965). On quantification of typology. *Linguistics 12,* 31–36.

Lamprecht, A. (1983). *Relationale Satzanalyse. Theorie und Praxis einer konsistenten Analyse englischer Satzstrukturen*. München: Hueber.

Nižníková, J., Sokolová, M. (1998). *Valenčný slovník slovenských slovies*. Prešov: Filozofická fakulta Prešovskej Univerzity.

Schumacher, H. (1986). *Verben in Feldern. Valenzwörterbuch zur Syntax und Semantik deutscher Verben*. Berlin, New York: de Gruyter.

Welke, K. (1988). *Einführung in die Valenz- und Kasustheorie*. Leipzig: Enzyklopädie.

2.10 VALENCE A SYNONYMIE

Hypotéza

Čím má sloveso větší valenci, tím má více synonym. Ověřte tuto hypotézu.

Postup

U mnohých sloves se vysoká valence může pojit s pronikáním jednoho slovesa do sémantické domény jiného slovesa, čímž může docházet k navyšování počtu jejich synonym. Připravte si opět tabulku (v elektronickém tabulkovém procesoru) obsahující valenci a počet synonym jednotlivých sloves. Následně vypočtěte nejprve korelaci a poté demonstrujte, že závislost sice tvoří přímku, ale ne horizontální. Jen málo valenčních slovníků uvádí také synonyma, a proto je každopádně lepší pracovat vedle valenčního slovníku také se slovníkem synonym.

Uveďte, zda daný vztah platí i ve světle analýzy několika dalších jazyků. Pokud ne, hypotézu upravte a doplňte vhodné hraniční podmínky.

Jelikož synonymie souvisí s polysémií, demonstrujte vliv obou proměnných (valence a polysémie) na synonymii a odvoďte příslušné vzorce.

Literatura

Srov. část 2.9

2.11 VALENCE A DÉLKA

Hypotéza

Čím má sloveso větší valenci, tím je kratší.

Postup

Frekvence „zkracuje" slova a současně dává slovesům možnost zvýšit jejich valenci. Z toho důvodu by mělo být možné pozorovat minimálně korelaci mezi délkou a valencí slovesa.

Připravte příslušná data a zkoumejte chování těchto dvou proměnných. Délku slova je třeba měřit v kanonické formě, tj. po lemmatizaci, a nikoli z hlediska počtu fonémů nebo grafémů, ale spíše z hlediska počtu slabik. Neočekáváme žádnou lineární relaci a nejsme schopni predikovat ani směr jejich vzájemné závislosti. Přijatelná je rovněž inverzní varianta závislosti, tzn. „čím je sloveso kratší, tím má větší valenci".

Literatura

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook.* Berlin, New York: de Gruyter, 760–774.

2.12 ŘÍDICÍ CYKLUS VALENCE

Problém

Určete valenci, synonymii, polysémii, frekvenci a délku součástí řídicího cyklu.

Postup

Připomeňte si notaci používanou v synergetické lingvistice a sestrojte odpovídající diagram. Určete požadavky a řídicí parametry uspořádání, které řídí procesy související s výslednou strukturou.

Literatura

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, 760–774.

2.13 VALENCE SUBSTANTIV A ADJEKTIV

Problém

Které ze vztahů mezi valencí a dalšími vlastnostmi, o nichž bylo pojednáno výše, platí rovněž pro substantiva a adjektiva?

Postup

Analogicky s výše uvedenými popisy shromážděte data týkající se nikoli sloves, ale substantiv a adjektiv. Proveďte analýzu příslušných výsledků a určete, které z těchto vztahů vypovídají o jasných tendencích a charakterizují rozdíly mezi danými třemi skupinami. Následně navrhněte hypotézu o vlastnostech slovních druhů, které určují formu jejich valenčních vztahů, a formulujte je jako hraniční podmínky.

Literatura

žádná

2.14 VALENCE: DISTRIBUCE VARIANT

Hypotéza

Distribuce počtu variant sloves odpovídá pozitivnímu negativně binomickému rozdělení.

Postup

Z valenčního slovníku nebo z internetového zdroje zjistěte u sloves počet jejich variant. Výběrový soubor by měl obsahovat co možná nejvíce sloves, optimálně celý slovník. Variantu slovesa lze operacionalizovat jako heslo ve valenčním slovníku s charakteristickým valenčním rámcem a většinou zvláštním významovým odstínem. Každé sloveso tak bude mít $x = 1, 2, \ldots$ variant, přičemž

$$P_x = \frac{\binom{k+x-1}{x}}{1-p^k} p^k q^x, \; x = 1, 2, \ldots$$

Ověřte tuto hypotézu, tj. aplikujte danou distribuci na zkoumaná data a proveďte chí-kvadrát test.

Hypotézu zdůvodněte nebo vytvořte jinou, pokud se distribuce s daty rozchází. Nalezněte důvody pro odmítnutí hypotézy, tj. věnujte pozornost derivaci negativně binomického rozdělení.

Literatura

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

2.15 VALENCE A VALENČNÍ RÁMEC

Hypotéza

Distribuce valenčních rámců sloves odpovídá Zipf-Mandelbrotovu rozdělení.

Postup

Pomocí vhodného valenčního slovníku zjistěte počet sloves, jejichž popis vypovídá o přítomnosti daného rámce. Příkladem valenčního rámce je

$$S_{nom} - PRED_{vf} - S_{dat}/\check{z}e\ SENT/S_{v\,ak},$$

jímž se popisuje valence slovesa s obligatorním subjektem a obligatorně přímým či nepřímým objektem, případně vedlejší větou (klauzí) uvozenou prostřednictvím *že* nebo předložkovým objektem s předložkou *v* (např. *věřit*: Já tomu věřím / Já mu věřím / Já věřím, že je to pravda / Já věřím v moc algebry). Možná jsou rovněž fakultativní (neobligatorní) doplnění.

Najdeme velké množství rámců, které se vyskytují pouze u jednoho slovesa (jeho varianty), o něco menší počet rámců, které vystihují chování dvou různých sloves atd. Rámec odpovídající nejvyššímu počtu sloves má rank 1. Výsledná ranková distribuce by se měla řídit Zipf-Mandelbrotovým rozdělením. Ověřte tuto hypotézu, tj. aplikujte danou distribuci na data a proveďte chí-kvadrát test.

Literatura

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

2.16 DISTRIBUCE SÉMANTICKÝCH SUBKATEGORIÍ ARGUMENTŮ

Hypotéza

Distribuce přípustných sémantických subkategorií argumentu ve valenčním rámci odpovídá pozitivnímu Poissonově rozdělení.

Postup

Vedle počtu a typu jednotlivých argumentů ve valenčním rámci poskytují některé valenční slovníky také informace o sémantických subkategoriích, což slouží k přesnějšímu výběru slov pro konkrétní typ argumentu. Mezi takové sémantické subkategorie patří např. +anim (možné jsosu pouze argumenty s rysem životnosti), -anim (argumenty s absencí rysu životnosti), podobně +/- lidské bytosti, +/- abstrakta, +/- kolektiva apod. Zjistěte počet argumentů s $x = 1, 2, \ldots$ možných subkategorií v daném valenčním rámci. Podíl každého slovesa na tomto čísle odpovídá počtu jeho komplementů, tzn. sloveso s dvěma komplementy představuje dvě samostatné položky. Proměnná x by se měla řídit distribucí

$$P_x = \frac{\lambda^x}{x!(e^\lambda - 1)} P_{x-1}, \;\; x = 1, 2, 3, \ldots$$

Ověřte tuto hypotézu, tj. aplikujte tuto distribuci na analyzovaná data, a proveďte chí-kvadrát test.

Literatura

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

2.17 POČET ARGUMENTŮ A POČET SÉMANTICKÝCH SUBKATEGORIÍ

Hypotéza

Počet přípustných sémantických subkategorií je lineárně funkčně závislý na počtu argumentů ve valenčním rámci.

Postup

Je evidentní, že čím více argumentů (komplementů) valenční rámec má, tím větší je počet přípustných sémantických subkategorií. Určete formu této závislosti. Lze předpokládat, že bude mít lineární podobu.

Ověřte tuto hypotézu, tj. aplikujte lineární funkci $y = ax + b$ na analyzovaná data a vypočtěte koeficient determinace.

Literatura

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

2.18 FREKVENCE A ALOMORFIE

Hypotéza

Počet alomorfů určitého morfému je rostoucí funkcí frekvence morfémů. Čím má nějaká jazyková jednotka větší frekvenci, tím je snadněji zapamatovatelná. Četnější jednotky proto mohou být nepravidelné, což lze opětovně využít k vytváření ekonomických, krátkých alomorfů. Dobrým příkladem lexémů s vysokou frekvencí je sloveso *být*, srov. množinu alomorfů jeho kořene {*by, buď, js, je*}, zatímco například spíše vzácný morfém *simul* se ve všech svých derivačních a inflekčních kontextech vyskytuje pouze v jediném tvaru (*simul-ovat, -ace, -tánní* apod.). Ověřte tuto hypotézu.

Postup

Na základě textových dat nebo dat z frekvenčního slovníku shromážděte datový korpus týkající se frekvence morfémů a počtu alomorfů příslušných morfémů. Pro každý morfém z vašeho výběrového souboru sestrojte bodový graf (frekvence morfémů, počet alomorfů) a opticky ověřte, zda jednotlivé body přibližně

tvoří plynulou přímku nebo alespoň vykazují nějakou zřejmou tendenci. Pokud to bude možné, aplikujte na tato data nějakou funkci.

Data vybírejte náhodně, nepracujte s morfémy vybranými selektivně. Jelikož počty alomorfů představují menší doménu než je frekvenční, hypotézu uplatněte obráceně a danou závislost uvádějte jako frekvence = f(počet alomorfů). Uvažujte samozřejmě průměrné frekvence.

Literatura
žádná

2.19 SÉMANTICKÁ RELEVANCE AFIXŮ (1)

Hypotéza
Čím má afix (kategorie) větší relevanci vzhledem k významu slova, tím je blíže k jeho kořeni. Ověřte tuto hypotézu.

Postup
Nejprve definujte nezávislé měřítko sémantické relevance, které by zohlednilo, do jaké míry se význam kompletního slovního tvaru liší od významu kořene. Poziční blízkost se definuje snáze: je dána počtem morfů mezi kořenem a zkoumaným afixem. Určete míru blízkosti a relevance u všech slov s afixy obsažených v textových datech a pokuste se nalézt odpovídající funkci.

Návrh způsobu měření sémantické relevance pojměte jako samostatný problém. Vypracujte škálovací postup. Stanovte empirickou distribuci míry relevance u všech tvarů slov v textu. Na základě teoretické argumentace definujte rozdělení pravděpodobnosti. Tento postup aplikujte na různé jazyky. Pomocí tohoto kritéria definujte rozdíl mezi jednotlivými jazyky.

Literatura

Bybee, J. L. (1985). *Morphology: a study of the relation between meaning and form*. Amsterdam, Philadelphia: J. Benjamins.

2.20 SÉMANTICKÁ RELEVANCE AFIXŮ (2)

Hypotéza

Čím větší je vnitřní sémantická koherence slova (tj. čím mají afixy vzhledem ke kořeni větší relevanci), tím větší je pravděpodobnost morfofonemického účinku. Ověřte tuto hypotézu.

Postup

Z výše definovaného (srov. část 2.19, „Sémantická relevance afixů [1]“) měřítka sémantické relevance odvoďte měřítko vnitřní sémantické koherence slova.

Určete míru koherence všech slov s afixy, která jsou obsažena v textových datech z nějakého flektivního nebo aglutinačního jazyka. U každého z těchto slov se zaměřte na případný morfofonemický účinek. Na své škále koherence si definujte vhodné intervaly a podle těchto intervalů pak slova roztřiďte do skupin. U každé z těchto skupin vypočtěte relativní frekvenci morfofonemických účinků

$$M_i = E_i / S_i ,$$

kde E_i je počet slov ve skupině *i* vykazujících morfofonemický účinek a S_i je počet všech slov ve skupině *i*. Skupiny mohou být nyní reprezentovány dvojicemi (C_i, M_i), kde C_i představuje průměrnou hodnotu soudržnosti slov ve skupině *i*. Prokažte v rámci datového korpusu existenci nějaké tendence nebo dokonce funkční závislosti, které by potvrzovaly danou hypotézu.

Literatura

Bybee, J. L. (1985). *Morphology: a study of the relation between meaning and form*. Amsterdam, Philadelphia: J. Benjamins.

2.21 SLOVOSLED A URČENÍ VÝCHODISKA VÝPOVĚDI

Hypotéza

Problematika kódování východiska výpovědi (tématu) již byla několikrát předmětem zkoumání a v tomto ohledu již bylo formulováno několik hypotéz. Různí se také názory na interpretace důvodů existence určitých struktur, jako je tomu v případě motivace k uspořádání slovosledných vzorců ve větě.

RÉMA> RÉMA-TÉMA> TÉMA-RÉMA> TÉMA (OPAKOVÁNÍ)

(nulové téma) (nulové réma)

Jedna možná interpretace vychází z míry kontinuity nebo prediktability (čím je téma evidentnější, tím menší je snaha o kódování nebo zdůrazňování), druhá z psycholingvistické zásady „zabývejte se nejprve tím, co nejvíce spěchá". Je zjevné, že do hry zde vstupují ekonomie, ikonicita a dekódování.

Ověřte hypotézu, že frekvence čtyř výše uvedených slovosledných struktur v textech se řídí rozdělením pravděpodobnosti, které patří do skupiny rozdělení používaných na modelování diverzifikace.

Postup

Na datech z textů pocházejících nejméně ze dvou jazyků určete frekvence výše uvedených čtyř slovosledných struktur a aplikujte na tato data příslušné rozdělení pravděpodobnosti. Získané výsledky interpretujte. Určete teoretický model mechanismu, kterým se řídí slovosledné struktury.

Literatura

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, 646–658.

Givón, T. (1985). Iconicity, isomorphism and non-arbitrary coding in syntax. In: Haiman, J. (ed.), *Iconicity in syntax*. Amsterdam, Philadelphia: J. Benjamins.

Wimmer, G., Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

2.22 SYNTAKTICKÉ VLASTNOSTI

Problém

Kvantitativní výzkum vlastností a vzájemných vztahů syntaktických struktur (srov. Köhler 1999; Köhler, Altmann 2000) byl doposud prováděn pouze na korpusech ze dvou jazyků (angličtiny a němčiny). Pokuste se tuto empirickou základnu rozšířit.

Postup

Najděte si syntakticky anotovaný korpus v nějakém jiném jazyce než němčině nebo angličtině, případně si takový korpus, byť malý, sestavte sami. Proveďte

analýzu korpusu způsobem, který bude analogický s výše uvedenými postupy, a porovnejte vaše výsledky s výsledky, ke kterým dospěli Köhler a Altmann.

Literatura

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1),* 46–57.

Köhler, R., Altmann, G. (2000). Probability dstributions of syntactic units and properties. *Journal of Quantitative Linguistics 7(1),* 189–200.

2.23 EFEKTIVITA SYSTÉMU SLOVNÍCH DRUHŮ

Problém

Vulanović (2008) se zabýval výpočtem teoretické efektivity různých typologicky osvědčených systémů slovních druhů. Zjistěte efektivitu existujících systémů slovních druhů. Zohledněte frekvenci odpovídajících jevů pozorovaných při komunikaci (v textech).

Postup

Způsob měření efektivity navržený Vulanovićem je založen na vlastnostech systému. Zohlednění frekvence výskytu slov patřících k určitému slovnímu druhu a četnost potřeby způsobu, jakým jsou v jazyce vyjadřovány propoziční funkce, může změnit obraz efektivity v reálném úzu.

Proveďte anotaci textů podle gramatických a lexikálních popisů některých jazyků z hlediska třídění na jednotlivé slovní druhy a z hlediska uplatňování propozičních funkcí. Formulujte metodiku měření efektivity, která by zohledňovala frekvenci, a aplikujte ji na anotované texty.

Literatura

Vulanović, R. (2008). A mathematical analysis of parts-of-speech systems. *Glottometrics 17,* 51–65.

2.24 DÉLKA A KOMPLEXITA SYNTAKTICKÝCH STRUKTUR

Problém

Kvantitativní zkoumání komplexity a délky syntaktických struktur, informačního obsahu, postavení v rámci vyššího konstituentu a dalších vlastností (srov. Köhler 1999; Köhler, Altmann 2000) je definováno a operacionalizováno s ohledem na gramatickou analýzu vět na základě frázových struktur. Definujte vlastnosti a vzájemné vztahy u vět analyzovaných podle dependenční gramatiky.

Postup

Najděte nejméně dvě vlastnosti, které lze přičíst částem stemmatu. Operacionalizujte je a odpovídajícím způsobem pak tyto vlastnosti změřte za použití syntakticky anotovaného textového korpusu, např. Pražského závislostního korpusu. Určete frekvenční distribuce jednotlivých vlastností a aplikujte na analyzovaná data teoretická rozdělení pravděpodobnosti. Budou-li tyto dvě (případně další) vlastnosti vykazovat funkční závislost, formulujte příslušnou hypotézu a aplikujte odpovídající funkci na zkoumaná data.

Literatura

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1),* 46–57.

Köhler, R., Altmann, G. (2000). Probability dstributions of syntactic units and properties. *Journal of Quantitative Linguistics 7(1),* 189–200.

2.25 GRAMATIKA, TEXT, KORPUS, JAZYK

Problém

Köhler (1999) a Köhler, Altmann (2000) vycházejí ve svých kvantitativně zaměřených studiích z analýzy vět z hlediska gramatiky frázových struktur. Gramatika užívaná konkrétně anotátory v rámci lancasterského projektu má navíc některé specifické vlastnosti. Do jaké míry jsou výsledky prezentované ve výše uvedených studiích popisem jednotlivých textů, potažmo korpusu, daného jazyka či vlastností zvolené gramatiky?

Postup

Seznamte se podrobně se syntaktickou analýzou, která vedla ke zpracování anotace v korpusu Susanne (Sampson 1995). Pokuste se zjistit, k jakým změnám ve výsledcích by došlo v závislosti na volbě konkrétních vlastností různých modifikací gramatických principů.

Literatura

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1),* 46–57.

Köhler, R., Altmann, G. (2000). Probability distributions of syntactic units and properties. *Journal of Quantitative Linguistics 7(1),* 189–200.

Sampson, G. (1995). *English for the Computer.* Oxford: Clarendon Press.

2.26 FUNKČNÍ ZÁVISLOSTI V SYNTAXI

Problém

Několik hypotéz týkajících se funkčních závislostí mezi syntaktickými vlastnostmi formuloval a empiricky ověřil Köhler (1999).

Některé z nich mají formu $y = Ax^b e^{cx}$:

- průměrná frekvence větného konstituentu jako funkce komplexity prvku,
- vztah mezi komplexitou a délkou,
- míra zanoření a její závislost na pozici větného konstituentu.

Tento vzorec teoreticky odvoďte nebo zdůvodněte ve vztahu k daným hypotézám.

Postup

Určete hypotetické mechanismy, které ovlivňují vazby mezi dvojicemi zkoumaných proměnných. Doporučujeme vycházet z modelovacího postupu popsaného v jedné z referenčních studií (Köhler 2006) a využít rozšířené verze nástrojů synergetické lingvistiky, o nichž je zde pojednáno.

Literatura

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1),* 46–57.

Köhler, R. (2006). Frequenz, Kontextualität und Länge von Wörtern – Eine Erweiterung des synergetisch-linguistischen Modells. In: Rapp, R., Sedlmeier, P., Zunker-Rapp, G. (eds.), *Perspectives on Cognition – A Festschrift for Manfred Wettler.* Lengerich: Pabst Science Publishers, 327–338.

2.27 ROZDĚLENÍ KOMPLEXITY

Hypotéza

Syntaktická komplexita klauzí vykazuje – stejně jako komplexita syntaktických konstrukcí obecně – hyper-Pascalovo rozdělení.

Postup

V níže uvedené studii (Köhler, Altmann 2000) bylo rozdělení komplexity syntaktických konstrukcí analyzováno a modelováno prostřednictvím hyper-Pascalova rozdělení. V rámci této práce tvořil předmětná data úhrn všech konstrukcí obsažených v korpusu. Teoretické úvahy, jež vedly ke koncipování daného modelu, by měly platit zejména pro jednotlivé druhy konstrukcí.

Shromážděte data týkající se komplexity, setřiďte je podle různých druhů konstrukcí, tj. na fráze a klauze, a ověřte danou hypotézu. Podaří-li se hypotézu na vašich datech potvrdit, ověřte, zda se parametry rozdělení liší v závislosti na druhu konstrukce.

Literatura

Köhler, R., Altmann, G. (2000). Probability distributions of syntactic units and properties. *Journal of Quantitative Linguistics 7(1),* 189–200.

2.28 INFORMAČNÍ STRUKTURA (1)

Hypotéza

Informační hodnota jednotlivých prvků syntaktické struktury klesá s jejich vyšším postavením v rámci dané struktury.

Postup

Informace (ve smyslu teorie informace) je změna v míře neurčitosti – nebo míra toho, kolik jednotek informací (bitů) je potřebných ke kódování. Informaci obsaženou v prvku syntaktické struktury lze měřit na základě počtu alternativ, které lze použít namísto daného prvku (paradigmatický vztah).

Danou hypotézu lze nejjednodušeji ověřit prostřednictvím syntakticky anotovaného korpusu. V nějakém dlouhém textu či korpusu spočítejte u každého typu struktury počet alternativ, které lze v určité pozici v rámci dané struktury použít: začněte s první pozicí v rámci určitého typu struktury (např. jmenné fráze) a určete počet typů jednotlivých prvků (typů frází a slovních tříd), kterou může daná konstrukce začínat (např. determinátor, vlastní jméno, zájmeno apod.). Poté pro jakýkoliv identifikovaný typ prvku určete počet jeho „následovníků", tj. typů prvků na pozici 2 atd. Logaritmus takového počtu vyjadřuje míru informační hodnoty prvku na dané pozici. Vezmete-li si logaritmus o dvojkovém základě, dostanete informaci vyjádřenou v bitech. Zkoumejte závislost míry informace na pozici za použití dat z co největšího množství různých jazyků.

Literatura

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1),* 46–57.

Köhler, R. (2000). A study on the informational content of sequences of syntactic units. In: Kuz'min, L. A. (ed.), *Jazyk, glagol, predloženie. K 70-letiju G. G. Sil'nitskogo.* Smolensk, 51–61.

2.29 INFORMAČNÍ STRUKTURA (2)

Problém

Zaměřte se na informační hodnotu jazykových prvků v paradigmatických vztazích.

Postup

Jednotlivé prvky či rysy, které utvářejí nějaké paradigma nebo kategorii, se nevyskytují se stejnou frekvencí. Informace spojené s určitým typem kategorie nebo prvku (srov. část 2.28, „Informační struktura [1]“) lze proto měřit pomocí konceptu entropie. Dochází tak k zohlednění rozdělení pravděpodobnosti daných prvků. Vypočítejte entropie prvků v syntaktické konstrukci na základě jejich frekvence v pozicích, které zaujímají v dané struktuře, a porovnejte získané poznatky s výsledky řešení předchozího problému. K jakým závěrům jste dospěli?

Literatura

Altmann, G. (1980). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1),* 46–57.

Köhler, R. (2000). A study on the informational content of sequences of syntactic units. In: Kuz'min, L. A. (ed.), *Jazyk, glagol, predloženie. K 70-letiju G. G. Sil'nitskogo. Smolensk,* 51–61.

2.30 DIVERZIFIKACE VIDU

Problém

V některých jazycích, např. slovanských, se slovesný vid (aspekt) vyjadřuje morfologickými prostředky. V jiných k tomu slouží různé analytické

konstrukty, fráze apod. Prokažte, že pokud tyto prostředky podléhají diverzifikačnímu procesu, tak je jejich frekvence ve vztahu s jejich délkou (srov. řešení tohoto problému v obecnější rovině v části 9.12, „Frekvence a produkční úsilí [pokračování]).

Postup

Za použití učebnice gramatiky daného jazyka zpracujte výčet všech forem vyjádření vidu. Následně se zaměřte na výskyt jednotlivých forem v rámci korpusu. Délku jednotlivých prostředků měřte různým způsobem, např. z hlediska počtu slov, počtu slabik, počtu morfémů apod. Nejprve vypočtěte korelaci mezi délkou a frekvencí, pak najděte funkci vyjadřující jejich vzájemnou závislost. Zakomponujte daný problém do systému synergetické lingvistiky.

Literatura

Bache, C. (1982). Aspect and Aktionsart: Towards a semantic distinction. *Journal of Linguistics 18(1),* 57–72.

Binnick, R. I. (1991). *Time and the verb: A guide to tense and aspect*. New York: Oxford University Press.

Binnick, R. I. (2006). Aspect and Aspectuality. In: Aarts, B., McMahon, M. S. (eds.), *The Handbook of English Linguistics*. Malden, MA: Blackwell Publishing, 244–268.

Comrie, B. (1976). Aspect: *An introduction to the study of verbal aspect and related problems*. Cambridge, New York: Cambridge University Press.

Dahl, Ö. (ed.) (2000). *Tense and Aspect in the Languages of Europe*. Berlin: de Gruyter.

Gautier, L., Haberkorn, D. (eds.) (2004). *Aspekt und Aktionsarten im heutigen Deutsch*. Tübingen: Stauffenburg.

Herweg, M. (1990). *Zeitaspekte*. Wiesbaden: Westdeutscher Verlag.

Hollebrandse, B., Hout, A. v., Vet, C. (eds.) (2005). *Crosslinguistic views on tense, aspect and modality*. Amsterdam: Rodopi, 317–401.

Chertkova, M.Y. (2004). *Vid or Aspect? On the Typology of a Slavic and Romance Category* [Using Russian and Spanish Material]. Vestnik Moskovskogo Universiteta, Filologia, 58(9–1), 97–122.

Kortmann, B. (1991). The Triad „Tense-Aspect-Aktionsart". *Belgian Journal of Linguistics 6,* 9–30.

Löbner, S. (2002). Is the German Perfekt a perfect perfect? In: Kaufmann, I., Stiebels, B. (eds.), *More than Words: A Festschrift for Dieter Wunderlich.* Berlin: Akademie Verlag, 369–391.

MacDonald, J. E. (2008). *The syntactic nature of inner aspect: A minimalist perspective*. Amsterdam, Philadelphia: J. Benjamins.

Richardson, K. (2007). *Case and aspect in Slavic*. Oxford, New York: Oxford University Press.

Sasse, H.-J. (2002). Recent activity in the theory of aspect: Accomplishments, achievements, or just non-progressive state? *Linguistic Typology 6(2),* 199–271.

Sasse, H.-J. (2006). Aspect and Aktionsart. In: Brown, E. K. (ed.), *Encyclopedia of language and linguistics* (*Vol. 1*). Boston: Elsevier, 535–538.

Smith, C. S. (1991). *The parameter of aspect*. Dordrecht, Boston: Kluwer Academic Publishers.

Tatevosov, S. (2002). The parameter of actionality. *Linguistic Typology 6(3)*.

Zalizniak, A. A., Šmelev, A. D. (2000). *Vvedenie v russkuju aspektologiju* [Úvod do ruské aspektologie]. Moskva: Jazyki russkoj kul'tury.

2.31 PÁDOVÁ HIERARCHIE

Problém

(1) Existuje nějaká hierarchie pádů?

(2) Existuje hierarchie pádových ukazatelů? Prostudujte tyto problémy a pokuste se formulovat příslušné hypotézy.

Postup

Zaměřte se na jazyk s dobře rozvinutým pádovým systémem, např. latinu, němčinu nebo nějaký slovanský jazyk. V textu určete počet jednotlivých pádů i výskyt jednotlivých koncovek. Lze k tomu využít „tagger" nebo tužku a papír. Výsledná tabulka bude vypadat takto:

TABULKA 2.31.1

	Nom	Gen	Dat	Acc	Voc	Loc	Instr	Abl	...
-a									
-ae									
-am									
-arum									
-as									
...									
nula									

Žádná další gramatická kategorie zde není relevantní.

(a) První problém vyřešte tím, že prokážete nerovnoměrnou distribuci pádů, což lze provést prostřednictvím testu homogenity součtů jednotlivých sloupců.

(b) Demonstrujte, že k pravidelnému rozdělení lze dospět, když se součty jednotlivých sloupců seřadí podle frekvence. Pokuste se toto rozdělení vysledovat empiricky (např. pomocí softwaru) a následně je zdůvodněte příhodnými argumenty. Například: nominativ se vyskytuje téměř ve všech větách, protože… Další pád se vyskytuje v menším množství vět, protože… atd. Zdůvodnění použité matematické formy podepřete argumenty založenými na principu proporcionality.

(c) Druhý problém vyřešte prostřednictvím testu homogenity součtů jednotlivých řádků.

(d) Určete rankovou distribuci koncovek a lingvisticky ji zdůvodněte.

(e) Lze konstatovat, že čím je koncovka kratší, tím má vyšší frekvenci?

(f) Věnujte pozornost diverzifikaci jednotlivých koncovek a řešte některé z problémů, které se s tímto jevem pojí. Prostudujte si části věnované problematice diverzifikace v této publikaci.

Literatura

Fan, F., Altmann, G. (2008). On meaning diversification in English. *Glottometrics 17,* 66–78.

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17,* 94–108.

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18,* 32–39.

Rothe, U. (ed.) (1991). *Diversification processes in language: grammar.* Hagen: Rottmann.

3 Sémantika

3.1 POLYSÉMIE SLOVES A SUBSTANTIV

Hypotéza

Podle Gentnera (1981) „vykazují obecná slovesa větší významovou šíři než obecná substantiva". Ověřte tuto hypotézu.

Postup

Z frekvenčního slovníku vyberte substantiva a slovesa různé míry frekvence. Ve výkladovém slovníku zjistěte, kolik mají významů, a vzájemně je porovnejte. Analýza by měla být provedena na materiálu z nějakého jiného jazyka, než je angličtina, neboť právě ta je nejčastěji předmětem podobných výzkumů. Vybírejte z jazyků bez explicitních (morfologických) znaků rozdílu mezi substantivem a slovesem. Platí daná hypotéza?

Nyní si pomocí výkladového slovníku sestavte systematické výběrové soubory substantiv a sloves, použijte například poslední sloveso a substantivum na každé stránce. Spočítejte jejich významy a porovnejte jejich průměry.

Z výkladového slovníku vyberte 100 substantiv, z nichž je možné odvozováním utvořit sloveso. Zjistěte počty významů substantiv a sloves a porovnejte je.

Zopakujte postup popsaný výše, ale tentokrát pracujte se slovesy, z nichž je možné odvodit substantiva.

Prostřednictvím téhož testu zkoumejte případ konverze v angličtině.

Zpracujte si seznam substantiv a sloves obsažených v určitém textu a prostřednictvím výkladového slovníku stanovte počet jejich významů. Vypočtěte průměrný počet významů u substantiv i sloves a vzájemně je porovnejte za použití t-testu. Jste schopni potvrdit danou hypotézu?

Tentýž postup uplatněte na různé texty. Je případný rozdíl mezi průměry způsoben daným typem textu? Pokud ano, formulujte novou hypotézu a ověřte ji na jiných jazycích.

Jste schopni identifikovat nějaké jiné faktory, které mohou mít vliv na diverzifikaci významu?

Jste schopni předložit nějaké psychologické, vývojové nebo epistemiologické zdůvodnění tohoto případného rozdílu?

Literatura

Gentner, D. (1981). Some interesting differences between verbs and nouns. *Cognition and Brain Theory 4(2),* 161–178.

Oguy, O. (2005). Aproksimativni metodi v semasiologičnich poslidženijach: rezuľtati ta perspektivi zastosuvanija. In: Altmann, G., Levickij, V., Perebijnis, V. (ed.), *Problemi kvantitativnoi lingvistiki – Problems of Quantitative Linguistics*. Černivci: RUTA, 134–148.

Wordnet: A lexical database for English. [online]. Dostupné z: http://wordnet.princeton.edu/

Germanet: A German Wordnet. [online]. Dostupné z: http://www.sfs.uni-tuebingen.de/GermaNet/

3.2 POLYSÉMIE SLOVNÍCH DRUHŮ

Problém

Různé slovní druhy mají tendenci vykazovat různé projevy polysémie. Ověřte tuto hypotézu.

Postup

Jedná se o zobecnění problému prezentovaného v části 3.1, „Polysémie sloves a substantiv".

Pomocí výkladového slovníku si prostřednictvím náhodného nebo systematického výběru sestavte soubor substantiv, sloves, adjektiv a adverbií.

Zjistěte si míru jejich polysémie (tj. počet významů) a u každého slovního druhu zvlášť určete jeho frekvenční distribuci. Rozdíl v polysémii je možné vyjádřit porovnáním průměrů distribucí nebo provedením testu homogenity.

Pokuste se ukázat, že menší šikmost distribuce se pojí s větší polysémií. K charakterizování polysémie využijte rovněž entropii a míru opakování (Strauss et al. 2014, kap. „Index opakování a entropie").

Porovnejte jednotlivé distribuce pomocí Ordova schématu (Strauss et al. 2014, kap. „Ordovo kritérium"). Znázorněte výsledky graficky a formulujte hypotézu ohledně vzniku polysémie u jednotlivých slovních druhů. Jste schopni předložit zdůvodnění případných rozdílů?

Sestavte si seznam slov z určitého textu roztříděných podle slovních druhů. Jednotlivým slovům přiřaďte míru polysémie na základě počtu jejich významů uváděných ve výkladovém slovníku. Následně u každého slovního druhu zvlášť určete distribuci významů (X = počet významů, Y = počet slov s x významy). Porovnejte jednotlivé distribuce pomocí Ordova schématu. Proveďte analýzu několika textů a na základě porovnání polohy jednotlivých slovních druhů v rámci Ordova schématu formulujte novou hypotézu.

Ověřte případný rozdíl mezi vzorky týchž slovních druhů pořízenými ze slovníku a z textu.

Literatura

Oguy, O. (2005). Aproksimativni metodi v semasiologičnich poslidženijach: rezuľtati ta perspektivi zastosuvanija. In: Altmann, G., Levickij, V., Perebijnis, V. (ed.), *Problemi kvantitativnoi lingvistiki – Problems of Quantitative Linguistics*. Černivci: RUTA, 134–148.

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

3.3 SYNONYMIE A MORFOLOGICKÁ PRODUKTIVITA

Hypotéza

Čím je slovo morfologicky produktivnější (tj. čím více odvozenin a složenin vytváří), tím je větší jeho synonymie. Ověřte tuto hypotézu.

Postup

Ze slovníku náhodně vyberte 100 slov a spočítejte, u kolika odvozenin a kompozit slouží dané slovo jako základ. Následně pomocí slovníku synonym zjistěte, kolik mají daná základová slova synonym. Prokažte, že vzájemná závislost mezi morfologickou produktivitou a synonymií, tj. $Syn = f(MP)$, je monotónně rostoucí funkce. Najděte vhodnou funkci a zdůvodněte její adekvátnost, případně zdůvodněte nezbytnost tohoto vztahu a zakomponujte obě proměnné do diferenciální rovnice, jejíž řešení bude vyjádřením dané závislosti.

Učiňte tuto závislost součástí synergetického řídicího cyklu a identifikujte další faktory ovlivňující morfologickou produktivitu, synonymii, nebo obojí.

Literatura

žádná

3.4 SYNONYMIE A POSTPOZIČNÍ FRÁZE

Hypotéza

Čím více postpozičních frází sloveso vytváří, tím větší je jeho synonymie. Ověřte tuto hypotézu.

Postup

Z anglického výkladového slovníku vyberte náhodně 50 sloves a k nim všechny postpoziční fráze, např. *get in, get out, get around, get off, get out of, get from under, get*

through atd. Poté ze slovníku synonym zjistěte, kolik má dané sloveso (v tomto případě *get*) synonym. Určete přímou závislost synonymie na počtu postpozičních frází, tj. zkoumejte vztah <počet různých postpozičních frází, počet synonym>, tj. *Syn = f(PF)*, najděte vhodnou funkci a zdůvodněte její adekvátnost.

Stejnou analýzu aplikujte pokud možno i na jiné jazyky a její výsledky porovnejte s výsledky získanými u angličtiny. Viz také část 2.9, „Valence a synonymie", a 3.3, „Synonymie a morfologická produktivita".

Bude-li výsledek pozitivní, zakomponujte jej do řídicího cyklu zmiňovaného v části 3.3.

Literatura

žádná

3.5 SÉMANTICKÉ ČLENĚNÍ PROSTORU

Problém

V každém jazyce dochází k sémantickému členění prostoru prostřednictvím různých slovních tříd a morfémů, např. prepozic, postpozic, prefixů, afixů nebo adverbií místa. Pro každou slovní třídu zvlášť určete příslušný sémantický prostor a vyhodnoťte některé z jeho vlastností.

Postup

Shromážděte všechna slova (morfémy) dané třídy představující prostor (místo, směr, přechodová fáze). Za použití definic a analytických prostředků obsažených v referenční literatuře níže definujte a vyhodnoťte některé z následujících vlastností: (1) u systému určení místa: přesnost, orientaci, symetrii, účinnost; (2) u systému určení směru: rozlišení, entropii těchto rozlišení, symetrii, synkretismus místa a směru.

Výsledky porovnejte s výsledky z jazyků analyzovaných v rámci studií uvedených v referenční literatuře. Výsledky zobecněte. Zaměřte se na jinou třídu jednotek, vypočtěte vlastnosti tohoto systému a porovnejte je s vlastnostmi třídy analyzované v prvním případě. Je prostor v obou třídách reprezentován shodně?

Proveďte kompletní analýzu nějakého jazyka (tj. všech tříd vyjadřujících místo, směr nebo přechodovou fázi) a formulujte „prostorovou filozofii" daného jazyka. Pokud to bude možné, zkombinujte vaše výsledky s psychologickými, etnohistorickými nebo geografickými východisky.

Literatura

Altmann, G., Dömötör, Z., Riška, A. (1968). The partition of space in Nimboran. *Beiträge zur Linguistik und Informationsverarbeitung 12,* 56–71.

Altmann, G., Dömötör, Z., Riška, A. (1968). Reprezentácia priestoru v systéme slovenských predložiek. *Jazykovedný časopis 19,* 25–40.

3.6 SYNONYMIE A MORFOLOGICKÝ STATUS SLOVA

Hypotéza

Čím je slovo morfologicky jednodušší, tím více má synonym. Ověřte tuto hypotézu.

Postup

Slova je možné na základě jejich morfologického statutu klasifikovat či rozřadit na jednoduchá, odvozená, reduplikovaná, složená, reduplikované složeniny, složeniny s derivací. Opatřete si náhodný výběrový soubor slov z výkladového slovníku a roztřiďte je do výše uvedených tříd, jež mohou být uspořádány na ordinální škále. Následně pomocí slovníku synonym zjistěte počet synonym každého slova v daném výčtu. U každé třídy určete průměrný počet synonym

a ověřte hypotézu, že počet synonym určitého slova je funkcí jeho morfologické komplexity, tj. *počet synonym = f (morfologická komplexita)*. Pokud daný trend nevykazuje lineární průběh, najděte vhodnou funkci a zdůvodněte ji. K zajištění synergetickolingvistických východisek využijte pojem specifikace (Köhler 2005).

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.

3.7 VÝZNAMY SLOVA (1)

Problém

Určete distribuci jednotlivých významů slov v textech.

Postup

Zpracujte si seznamy slov z několika textů tvořících lexikálně zjednoznačněný (desambiguovaný) korpus, který obsahuje anotace jednotlivých slovních významů. Určete rozdělení pravděpodobnosti těchto významů ve vztahu ke každému slovu a každému textu.

(a) Potvrzují vaše zjištění závěry uváděné v literatuře o jazykové diverzifikaci (srov. seznam referenční literatury)?

(b) Lze najít vzájemné vztahy mezi vámi zjištěnou distribucí a
   i. délkou textu,
   ii. typem textu,
   iii. slovním druhem,
   iv. frekvencí slova,
   v. stářím slova nebo
   vi. délkou slova?

Zjistíte-li nějakou vzájemnou souvislost, vyjádřete ji prostřednictvím prosté funkce. Porovnejte své výsledky s výsledky z části 3.2, „Polysémie slovních druhů", které se týkaly slovníkových dat.

Literatura

Best, K.-H. (2009). Diversifikation des Phonems /r/ im Deutschen. *Glottometrics 18,* 26–31.

Laufer, J., Nemcová, E. (2009). Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18,* 13–35.

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18,* 32–39.

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17,* 97–111.

Rothe, U. (ed.) (1991). *Diversification process in language: grammar.* Hagen: Rottmann.

Sanada, H. (2009). *Diversification of postpositions in Japanese*.

3.8 VÝZNAMY SLOVA (2)

Problém

Určete počet jednotlivých významů slov realizovaných v textech a tematických doménách.

Postup

Z nějakého sémanticky anotovaného textového korpusu shromážděte data týkající se polysémie slov, tj. počet jednotlivých významů slov použitých v analyzovaných textech (na rozdíl od polysémie udávané slovníkem). Určete distribuce frekvence a pravděpodobnosti, pokud najdete dostatečně nejednoznačná slova. Porovnejte svá zjištění s polysémií slov udávanou ve vztahu k jednotlivým

textům nebo tematickým doménám. V případě nesouladu tuto skutečnost zdůvodněte.

Literatura

Arapov, M. V. (1987). Upotrebiteľnosť i mnogoznačnosť slova. *Učenye Zapiski Tartuskogo Universiteta 774*, 15–28.

Levickij, V. (2005). Polysemy. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, 456–464.

3.9 DISTRIBUCE SYNONYMIE SLOV

Problém

Určete distribuci synonymie slov.

Postup

Pomocí nějakého slovníku sestavte náhodný výběrový soubor slov a tato slova následně vyhledejte ve slovníku synonym nebo použijte Wordnet, Germanet či jiný elektronický zdroj (v závislosti na zkoumaném jazyku a možnostech). U každého slova zjistěte počet synonym a tato čísla uspořádejte podle velikosti.

Lze očekávat, že zkoumaným datům bude odpovídat nějaké specifické rozdělení pravděpodobnosti? Jaké konkrétně? Nejprve určete empirické rozdělení (nebo prostou funkci) a proveďte test shody a test dobré shody. Bude-li výsledek pozitivní, funkci zdůvodněte, sestavte odpovídající diferenciální nebo diferenční rovnici a formulujte příslušnou hypotézu. Jedná se o obvyklé induktivně-deduktivní pojetí, které se často využívá při koncipování teorií.

Literatura

Wimmer, G., Altmann, G. (2001). Two hypotheses on synonymy. In: Ondrejovič, S., Považaj, N. (eds.), *Lexicographica 1999*. Bratislava: Veda, 218–225.

Wordnet: A lexical database for English. [online]. Dostupné z: http://wordnet.princeton.edu/

Germanet: A German Wordnet. [online]. Dostupné z: http://www.sfs.uni-tuebingen.de/GermaNet/

3.10 SYNONYMIE A POLYSÉMIE

Problém

Najděte vzájemnou souvislost mezi synonymií a polysémií slov.

Postup

Za použití výkladového slovníku sestavte náhodný výběrový soubor slov náležejících k témuž slovnímu druhu a zjistěte počet jejich významů (polysémii). Řiďte se způsobem značení polysémie ve slovníku. Poté zjistěte ve slovníku synonym nebo na Wordnetu počet synonym každého z těchto slov. Následně prokažte, že průměrný počet synonym je mocninná funkce počtu významů. Tento vztah vychází z Köhlerova synergetického řídicího cyklu (1986, 2005) a jeho existenci se podařilo potvrdit pouze v jednom případě, u italštiny.

Věnujte pozornost dalším jazykům a pokuste se pro tuto hypotézu získat obecnější platnost.

Literatura

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

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch*. Berlin, New York: de Gruyter, 760–774.

Ziegler, A., Altmann, G. (2001). Beziehung zwischen Synonymie und Polysemie. In: Ondrejovič, S., Považaj, M. (eds.), *Lexicographica 1999.* Bratislava: Veda, 226–229.

3.11 SYNONYMIE, DÉLKA A FREKVENCE SLOV

Problém

Najděte vzájemnou souvislost mezi synonymií a délkou slov a synonymií a frekvencí slov.

Postup

Nejprve navrhněte hypotézy o možné souvislosti mezi (a) frekvencí slova a počtem jeho synonym a (b) délkou slova a počtem jeho synonym. Formulujte tyto hypotézy ve smyslu matematických funkcí. Poté za využití slovníku vytvořte náhodný výběrový soubor slov a určete jejich délku. Frekvence slov je možné zjistit z textového korpusu. Synonymii je možné zjistit buď ze slovníku synonym nebo automaticky z nějakého elektronické zdroje (Wordnet, Germanet apod.). Aplikujte funkce na zkoumaná data a na základě výpočtu determinačních koeficientů R^2 vyhodnoťte míru dobré shody.

Proveďte interpretaci výsledků. Snažte se získané výsledky včlenit do řídicího cyklu předchozích problémů.

Literatura

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch*. Berlin, New York: de Gruyter, 760–774.

Wimmer, G., Altmann, G. (2001). Two hypotheses on synonymy. In: Ondřejivič, S., Považaj, M. (eds.), *Lexicographica 1999.* Bratislava: Veda, 218–225.

4 Lexikologie

4.1 DEFINIČNÍ ŘETĚZCE (U SLOVES A ADJEKTIV)

Hypotéza

Slovesa a adjektiva mají kratší hyperonymní řetězce než substantiva. Ověřte tuto hypotézu.

Postup

V prvním svazku knihy *Kvantitativní lingvistika. Vybrané problémy 1* (Strauss et al. 2014, kap. „Lexikální řetězce"), bylo jedním z úkolů vytvořit hyperonymní řetězce substantiv a změřit délku těchto řetězců. Tímto způsobem byla zároveň stanovena frekvenční distribuce. Pomocí výkladového slovníku sestavte výběrový soubor čítající 100 sloves a 100 adjektiv. V obou případech určete distribuce délky hyperonymních řetězců a u každého zvlášť ověřte, zda

(a) jsou větve distribucí stejně dlouhé jako u substantiv. Základní data o německých a polských substantivech lze najít u Samborové a Hammerla (1991).

(b) Najděte model formy distribucí délky řetězců. Vyjděte z předpokladu, že čím je řetězec delší, tím menší je pravděpodobnost jeho doplnění o další hyperonymum, protože každé další (obecnější) hyperonymum je spíše součástí odborné slovní zásoby a tudíž zvyšuje nároky na kódování a paměť (srov. Köhler 2005). Při svém odvozování vycházejte pokud možno z Wimmer-Altmannovy sjednocené teorie.

(c) Zaměřte se blíže na důvody různé délky definičních řetězců u substantiv, sloves a adjektiv. Předložte nejen lingvistické zdůvodnění, ale využijte také poznatků z jiných vědních oborů (např. biologie nebo fyziky).

Literatura

Ballmer, T. T., Brennenstuhl, W. (1986). *Deutsche Verben: eine sprachanalytische Untersuchung des deutschen Verbwortschatzes*. Tübingen: Narr.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, 760–774.

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

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, 648–659.

4.2 DIACHRONNÍ STABILITA SLOVNÍCH TŘÍD

Problém

Zástupci některých slovních druhů (např. zájmen) vykazují větší intenzitu diachronní stability než je tomu u jiných slovních druhů. Za použití dat z románských jazyků publikovaných Kapitanem (1994) stanovte (1) indikátor intenzity diachronní stability a (2) indikátor homogenity diachronní stability.

Postup

Data Kapitana jsou ve zkrácené formě uvedena v tabulce 4.2.1.

TABULKA 4.2.1, Diachronní stabilita slovních druhů v románských jazycích (prvních 1 000) (Kapitan 1994)

Slovní druhy	Latina	Rumunština	Italština	Francouzština	Španělština	Portugalština
S	355	102	196	167	177	183
V	323	85	136	108	138	138
ADJ	178	49	84	70	77	79
ADV	87	11	19	15	19	18
NUM	5	5	4	4	5	5
PRON	6	5	5	4	5	5
PREP	13	7	6	7	8	8
KONJ	26	6	7	6	5	5
INTERJ	3	1	1	1	1	1

Intenzita diachronní stability je dána průměrem „zděděných" slov, homogenita diachronní stability je dána mírou rovnocennosti podílů diachronní stability. Navrhněte nějaké indikátory, pokud možno včetně jejich rozptylu, a určete pořadí jednotlivých slovních druhů podle (1) a (2). Zkoumejte i další jazykové rodiny. Vyskytují-li se v nich jiné slovní druhy, danou klasifikaci upravte. Proveďte interpretaci všech výsledků a zdůvodněte je.

Literatura

Kapitan, M. E. (1994). Influence of various system features of Romance words on their survival. *Journal of Quantitative Linguistics 1(*3), 237–250.

4.3 FREKVENCE A DIACHRONNÍ STABILITA SLOV

Hypotéza

„Čím je slovo frekventovanější, tím větší má šanci se v jazyce udržet." (Kapitan 1994: 242). Ověřte tuto hypotézu.

Postup

Potřebná data nejsnáze získáme analýzou textů ze dvou historicky odlišných fází téhož jazyka. Jinou možností je rozbor textů v nějakém klasickém jazyce a jazycích, které se z něj vyvinuly, např. v latině a nějakém současném románském jazyce. S daty druhého typu pracoval Kapitan (1994), který kódoval dané frekvence v intervalech $(2^n, 2^{n+1})$, nejfrekventovanější slova pak v intervalu $<2^0, 2^5>$. V tabulce 4.3.1 tak vidíme intervaly rankového pořadí, které lze vyjádřit jako řadové číslovky. V tabulce jsou uvedeny počty slov, které se v rámci jednotlivých frekvenčních skupin zachovaly v pěti různých románských jazycích.

TABULKA 4.3.1, Počty latinských slov podle jednotlivých frekvenčních intervalů do 1 000, které se zachovaly v pěti románských jazycích
(Kapitan 1994: 242)

Frekvenční interval	Latina	Rumunština	Italština	Francouzština	Španělština	Portugalština
<1–32> 1	32	21	25	22	25	25
<33–64> 2	32	18	18	18	17	17
<65–128> 3	64	33	45	38	41	41
<129–256> 4	128	42	64	60	65	64
<257–512> 5	256	71	132	103	119	119
<513–995> 6	483	88	175	142	169	177

(1) Najděte funkci charakterizující pokles *relativního* počtu slov, která se v jazyce zachovala, a pokuste se odůvodnit její formu. Věnujte pozornost třetí skupině, která se odchyluje od klesajícího trendu.

(2) Zaměřte se podrobněji na homogenitu slov, která se v daných pěti jazycích zachovala.

(3) Dospějeme ke stejné funkci, pokud si určíme odlišné frekvenční intervaly? Zobecněte daný problém.

Literatura

Arapov, M. V., Cherc, M. M. (1974). *Matematičeskie metody v istoričeskoj lingvistike*. Moskva: Nauka. [*Matematische Methoden in der historischen Linguistik*. Bochum: Brockmeyer,1983.]

Kapitan, M. E. (1994). Influence of various system features of Romance words on their survival. *Journal of Quantitative Linguistics 1(3)*, 237–250.

4.4 DISTRIBUCE SLOVNÍCH TŘÍD 2[1]

Hypotéza

Slovní třídy (slovní druhy atd.) vykazují pravidelnou rankovou sekvenci. Ověřte, zda se data dají modelovat nějakou funkcí, a zobecněte ji na jakýkoli druh slovních tříd.

Postup

V prvním svazku knihy *Kvantitativní lingvistika. Vybrané problémy 1* (Strauss et al. 2014, kap. „Distribuce slovních tříd") byla ranková sekvence prezentována jako distribuce, zatímco zde navrhujeme hledat „nejlepší" prostou nebo

1 Srov. tentýž problém v prvním svazku knihy *Kvantitativní lingvistika. Vybrané problémy 1* (Strauss et al. 2014, kap. „Distribuce slovních tříd"). Zde bude provedeno jeho zobecnění.

pravděpodobnostní funkci vyjadřující pravidelné zastoupení slovních tříd. Nejjednodušší je uvažovat slovní druhy. První problém spočívá ve shromáždění všech publikovaných dat. Slovní druhy viz Best (1994, 1997, 2000, 2001); Hammerl (1990); Judt (1995); Schweers, Zhu (1991); Tuzzi, Popescu, Altmann (2009); Zhu, Best (1992); Ziegler (1998, 2001); inflekční třídy viz Belonogov (1964); tvary sloves viz Bull (1947) a Robbins (1926); časy viz Hills, Anderson (1929, 1930); osobní zájmena Hills, Anderson (1931); afixy Pierce (1961, 1962); Veenker (1968, 1969, 1973, 1975, 1976) atd. K tomuto tématu existuje značné množství literatury.

Najděte nejlepší funkci pro všechny třídy. Pokud to bude nutné, použijte modifikovanou nebo zobecněnou verzi některé funkce. Běžně se používají tyto hlavní funkce:

Zipfova funkce zeta: $$f(r) = \frac{C}{r^a},\ r = 1, 2, 3, \ldots$$

Mandelbrotova funkce: $$f(r) = \frac{C}{(r+a)^b},\ r = 1, 2, 3, \ldots$$

Zipf-Aleksejevova funkce: $$f(r) = ar^{-b-c\ln r}\ , r = 1, 2, 3, \ldots$$

Altmannova funkce: $$f(r) = \frac{\binom{b+r}{r-1}}{\binom{a+r}{r-1}} f(1),\ r = 1, 2, 3, \ldots$$

Negativně hypergeometrické rozdělení:

$$P(r) = \frac{\binom{M-r-2}{r-2}\binom{K-M+n-r}{n-r+1}}{\binom{K+n-1}{n}},\ r = 1, 2, \ldots, n+1$$

Testovat lze i různé další funkce sloužící jako modely frekvence slov a fonémů/písmen. Najděte nejlepší empirický výsledek, lingvisticky odůvodněte danou funkci a propojte ji s Wimmer-Altmannovou (2005) obecnou teorií.

Uvažujte první čtyři funkce jako prosté (tj. nenormalizované) sekvence, poslední představuje pravidelné rozdělení.

Literatura

Belonogov, G. G. (1964). Raspredelenie častot pojavlenija flektivnych klassov russkich slov. *Problemy kibernetiki 11*, 189–198.

Best, K.-H. (1994). Word class frequencies in contemporary German short prose texts. *Journal of Quantitative Linguistics 1(2)*, 144–147.

Best, K.-H. (1994). Zur Wortartenhäufigkeit in Texten deutscher Kurzprosa der Gegenwart. *Glottometrika 16*, 276–285.

Best, K.-H. (2000). Verteilungen der Wortarten in Anzeigen. *Göttinger Beiträge zur Sprachwissenschaft 4*, 37–51.

Best, K.-H. (2001). Zur Gesetzmäßigkeit der Wortartenverteilungen in deutschen Pressetexten. *Glottometrics 1*, 1–26.

Bull, W. E. (1947). Modern Spanish verb-form frequencies. *Hispania 30,* 451–466.

Hammerl, R. (1990). Untersuchung zur Verteilung der Wortarten im Text. *Glottometrika 11*, 142–156.

Hills, E. C., Anderson, J. O. (1929). The frequency of moods and tenses of verbs in recent Spanish plays. *Hispania 12*, 604–606.

Hills, E. C., Anderson, J. O. (1930). The frequency of verbs and tenses in recent Spanish plays. *Hispania 13*, 413–416.

Hills, E. C., Anderson, J. O. (1931). The relative frequency of Spanish personal pronouns. *Hispania 14,* 335–337.

Judt, B. (1995). *Wortartenhäufigkeiten im Deutschen und Französischen*. Göttingen: Staatsexamensarbeit.

Pierce, J. E. (1961). A frequency count of Turkish affixes. *Anthropological Linguistics 3*, 31–42.

Pierce, J. E. (1962). Frequencies of occurrence of affixes in French. *Anthropological Linguistics* 6, 30–41.

Robbins, F. E. (1926). Statistics of Greek verb-forms. *Classical Journal 15,* 101–108.

Schweers, A., Zhu, J. (1991). Wortartenklassifizierung in Lateinischen, Deutschen und Chinesischen. In: Rothe, U. (ed.), *Diversification processes in language: grammar.* Hagen: Rottmann, 157–165.

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1.* Olomouc: Univerzita Palackého v Olomouci.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Parts-of-speech diversification in Italian texts. *Glottometrics 19.*

Veenker, W. (1975). *Verzeichnis der ungarischen Suffixe und Suffixkombinationen*. Hamburg: Societas Uralo-Altaica.

Veenker, W. (1969). *Vogul suffixes and pronouns. An index a tergo*. The Hague: de Gruyter.

Veenker, W. (1973). *Verzeichnis der ostostjakischen (Vach) Suffixe und Suffixkombinationen (unter Einschluß der wichtigsten Pronomina)*. Hamburg: Societas Uralo-Altaica.

Veenker, W. (1975). *Verzeichnis der čeremissischen Suffixe und Suffixkombinationen*. Hamburg: Finnisch-Ugrisches Seminar.

Veenker, W. (1976). *Verzeichnis der votjakischen Suffixe und Suffixkombinationen*. Hamburg: Finnisch-Ugrisches Seminar.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, 791–907.

Zhu, J., Best, K.-H. (1992). Zum Wort im modernen Chinesisch. *Oriens Extremus 35,* 45–60.

Ziegler, A. (1998). Word class frequencies in Brazilian-Portuguese press texts. *Journal of Quantitative Linguistics 5,* 269–280.

Ziegler, A. (2001). Word class frequencies in Portuguese press texts. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček.* Trier: Wissenschaftlicher Verlag, 295–312.

4.5 POROVNÁVÁNÍ SLOVNÍ ZÁSOBY

Problém

Proveďte vyhodnocení všech dostupných metod používaných při porovnávání slovní zásoby dvou textů.

Postup

První část tohoto úkolu spočívá v dosti mechanické práci. V Köhlerově Bibliografii kvantitativní lingvistiky (1995) si najděte všechny práce, které se o tomto tématu zmiňují. Druhá fáze bude obnášet prohledávání internetu na základě různých klíčových slov (přisuzování autorství, mezitextová vzdálenost atd.). Další odkazy lze najít i v některých jiných pracích z poslední doby, např. Rudman (1998), Merriam (2003), Labbé (2007) atd.

Jakmile budete mít k dispozici přehled metod a vzorců, zaměřte se blíže na matematické vlastnosti navrhovaných vzorců, tj. určete například oblasti koeficientů, odvoďte jejich rozptyly a zkoumejte jejich chování v závislosti na rostoucí velikosti výběrového souboru, zejména budete-li pracovat s asymptotickými kvantitami.

Prostřednictvím všech shromážděných indikátorů podobnosti porovnejte nějaké dva texty. Vyhodnoťte efektivitu těchto indikátorů. Budete-li srovnávat více než dva texty, neprovádějte žádné klasifikace ani analýzy vedoucí k určení možného autorství. Máte-li k dispozici srovnatelné texty vzniklé v určitém časovém sledu, např. přednášky autorů při přebírání Nobelovy ceny za literaturu, novoroční projevy prezidentů apod., demonstrujte na nich, že podobnost se pojí s časovým odstupem. Kterým indexem je možné nejlépe vyjádřit tuto závislost?

Proveďte analýzu všech stávajících způsobů argumentace ohledně přisuzování autorství, identifikujte jejich slabiny, proveďte rekapitulaci dosavadních kritik a pokuste se tuto oblast systematizovat.

Literatura

Brunet, E. (1988). Une mesure de la distance intertextuelle: la connexion lexicale. Le nombre et le texte. *Revue informatique et statistique dans les sciences humaines 24(1–4)*, 81–116.

Klavina, S. P. (1977). *Sopostavlenie funkcionaľnych stilej latyšskogo jazyka (lingvostatističeskoe issledovanie)*. Vilnius, Diss.

Köhler, R. (1995). *Bibliography of quantitative linguistics*. Amsterdam, Philadelphia: J. Benjamins.

Labbé, C. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics 14(1),* 33–80.

Labbé, C., Labbé, D. (2001). Inter-textual distance and authorship attribution Corneille and Moliere. *Journal of Quantitative Linguistics 8,* 212–231.

Merriam, T. (2003). An application of authorship attribution by intertextual distance in English. *Corpus 2,* 167–182.

Muller, Ch. (1968). *Initiation a la statistique linguistisque*. Paris: Larousse.

Muller, Ch. (1977). *Principes et méthods de statistique lexicale*. Paris: Hachette université.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities,* 351–365.

Tuldava, J. (1971). Statističeskij metod sravnenija leksičeskogo sostava dvuch tekstov. *Linguistica 4,* 199–220.

Tuldava, J. (1998). *Probleme und Methoden der quantitative-systemischen Lexikologie*. Trier: Wissenschaftlicher Verlag.

Viprey, J.-M., Ledoux, C. N. (2006). About Labbé´s "Intertextual Distance". *Journal of Quantitative Linguistics 13(2–3),* 265–283.

4.6 OBVYKLOST SLOV

Problém

V tradici české korpusové lingvistiky se pracuje s pojmem obvyklost slov (srov. zejména Savický, Hlaváčová 2002). Prokažte, že se vlastně jedná o jiným způsobem operacionalizovanou polytextualitu.

Postup

Savický a Hlaváčová považují korpus za souvislou sekvenci slov. Tuto sekvenci pak dále dělí na stejně dlouhé segmenty, přičemž pozorují výskyt daného slova v těchto segmentech. Přesně takové je i východisko zákona Frumkinové (srov. Strauss et al. 2014, kap. 9, a seznam referenční literatury tamtéž), nicméně v tomto případě dochází k jeho zobecnění z jednoho textu na celý korpus. V této nové formě však ještě lépe odpovídá konceptu polytextuality. Tito autoři měří obvyklost slov pomocí empirického průměru výskytů (*average reduced frequency*, *ARF*). Prokažte, že takto získaná hodnota je identická s předpokládanou střední hodnotou negativně hypergeometrického rozdělení, které reprezentuje Frumkinové zákon.

Dalším měřítkem obvyklosti, které zavedli Savický a Hlaváčová (2002), je průměrná logaritmická vzdálenost (*average logarithmic distance*, *ALD*)

$$ALD = \frac{1}{N} \sum_{i=1}^{f} d_i \log_{10} d_i \ ,$$

kde d_i jsou vzdálenosti mezi výskytem jednoho a téhož slova, f je frekvence slova a N je součet vzdáleností. Považujte d_i za geometricky distribuovanou proměnnou a odvoďte předpokládanou průměrnou logaritmickou vzdálenost (*ALD*).

Literatura

Altmann, G., Burdinski, V. (1982). Toward a law of word repetitions in text-blocks. *Glottometrika 4,* 147–167.

Frumkina, R. M. (1962). O zakonach raspredelenija slov i klassov slov. In: Mološnaja, T. N. (ed.), *Strukturno-tipologičeskie issledovanija.* Moskva: ANSSSR, 124–133.

Savický, P., Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics 9(3),* 215–231.

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

4.7 INDIKÁTOR ASOCIACE

Problém

V gramatice, lexikologii a textologii se často užívá asociační index vycházející z teorie informace, tj.

$$I(w_1, w_2) = \log\left(\frac{N \times f(w_1, w_2)}{f(w_1) \times f(w_2)}\right),$$

kde w_1 a w_2 jsou dvě různá slova (nebo jiné jednotky), $f(w_i)$ (i = 1, 2) je frekvence slov v korpusu o velikosti N (ve vzorci lze vynechat) a $f(w_1, w_2)$ je společný výskyt těchto slov. Odvoďte rozptyl tohoto indikátoru.

Postup

Pokud budete uvažovat N, tak nejprve přesně definujte, zda je N počet slov nebo vět v textu. Následně definujte termín *spoluvýskyt*: chystáte se zkoumat bezprostřední sousedství nebo společný výskyt ve větě – nepřímé nebo přímé sousedství? Poté odvoďte rozptyl za použití metody Taylorova rozvoje.

Na základě rozptylu vytvořte asymptotický test významnosti indikátoru a otestujte všechny asociace jednoho slova. Seřaďte je podle síly asociace (kvantilu normálního rozdělení) a určete hranici, na níž lze dva spolu se vyskytující prvky považovat za kompozitum.

Literatura

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

Church, K., Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics 16*, 22–29.

Cramér, H. (1946). *Mathematical methods in statistics*. Princeton: Princeton University Press.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19,* 61–74.

Ferrer i Cancho, R., Reina, F. (2002). Quantifying the semantic contribution of particles. *Journal of Quantitative Linguistics 9(1),* 35–47.

Han, D., Ito, T., Furugori, T. (2002). Structural analysis of compound words in Japanese using semantic dependency relations. *Journal of Quantitative Linguistics 9(1),* 1–17.

Han, D., Kato, K., Furugori, T. (2001). *Automatická segmentace složených slov v japonštině na základě kontextových informací.* [japonsky]. Odborná zpráva IEICE, NLC 2001–05, 29–34.

Hurt, J. (1976). Asymptotic expansion of functions in statistics. *Aplikace Matematiky 21,* 444–456.

Kabayashi, Y., Tolunaga, T., Tanaka, H. (1994). Analysis of Japanese compound nouns using collocational information. In: *Proceedings of the 15th International Conference on Computational Linguistics (COLING´94).* Kyoto, 865–869.

Kendall, M. G., Stuart, A. (1969). *The advanced theory of statistics I, II*. London: Griffin.

Li, W. (1989). *Mutual information functions of natural language texts*. Santa Fe Institute Working Papers.

Manning, Chr. D., Schütze, H. (1999). *Foundations of statistical naural language processing.* Cambridge, London: The MIT Press.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician 46(1),* 27–29.

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

4.8 STABILITA SLOV

Problém

K obvyklosti slov (srov. část 4.6, „Obvyklost slov") má blízko pojem statistické stability slov. Hlavní rozdíl je v tom, že obvyklost slov vychází z úseků textu stejné délky, zatímco stabilita slov se týká odstavců. Vypočtěte stabilitu slov v jednom určitém textu a v celém korpusu.

Postup

Koeficient statistické stability slov zavedl Marusenko (1983) ve formě

$$WS = \frac{F_w \cdot m_w}{N \cdot n},$$

kde

F_w = frekvence daného slova w v textu
m_w = počet odstavců obsahujících slovo w
N = počet slov v textu
n = počet odstavců v textu.

Tento koeficient lze využít při odhadování důležitosti slova (srov. Zubov 2004) a obsahu textu.

Určete interval stability slov (WS), odvoďte jeho předpokládanou hodnotu a v každém případě jeho rozptyl. Uvažujte N a n jako konstantní charakteristiky daného textu, tj. musíte provést odhad distribuce F_w a m_w.

I pokud se nepodaří najít příslušné teoretické východisko, proveďte následující analýzy: (a) Vypočtěte *WS* pro každé slovo textu a slova seřaďte od nejvyšší po nejnižší *WS*. (b) Formulujte tvrzení o pozicích jednotlivých slovních druhů. (c) Proveďte teoretické odvození distribuce *WS* a porovnejte ji se svými empirickými výsledky. (d) Porovnejte *WS* s jinými koeficienty tohoto druhu.

Jaký je rozdíl mezi stabilitou slov a zákonem Frumkinové?

Literatura

Abracos, J., Lopes, G. P. (1997). Statistical methods for retrieving most significant paragraphs in newspaper articles. In: *ACL/EACL Workshop on Intelligent Scalable Text Summarization*, 51–57.

Akišina, O. V. (2001). Formaľnoje vyraženie osnovnogo soderžanija anglojazyčnogo reklamnogo teksta. In: Zubov, A. V. (ed.), *Materialy ežegodnoj naučnoj konferencii studentov i magistrantov universiteta, 5–6 aprelja 2000g. Časť vtoraja*. Minsk: MGLU, 3–8.

Alavi Džafar, A. (2000). Formaľnoe predstavlenie osnovnogo soderžanija anglijskich korotkich raskazov. In: Zubov, A. V. (ed.), *Materialy ežegodnoj naučnoj konferencii prepodavaelej i aspirantov universiteta, 5–6 aprelja 2000. Časť treťja*. Minsk: MGLU, 3–8.

Bolšakova, J. G. (2000). Statistika v ocenke osnovnogo soderžanija tekstov-opisanij finansovych operacij. In: Zubov, A. V. (ed.), Materialy ežegodnoj naučnoj konferencii prepodavatelej i aspirantov universiteta, 5–6 aprelja 2000. Časť vtoraja. Minsk: MGLU, 6–7.

Brandow, R., Mitze, K., Rau, L. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing and Management 31(5)*, 675–685.

Čaplja, A. I., Čaplja, S. G., Zubov, A. V. (1973). Avtomatičeskij otbor ključevych i poliključevych slov. In: Čaplja, A. I. (ed.), *Sbornik naučnych soobščenij fakulteta inostrannych jazykov*. Machačkala: DGU, 76–93.

Marusenko, M. A. (1983). O formirovanii slovnika slovarja statističeski ustojčivych naučno-techničeskich terminov. In: *Strukturnaja i prikladnaja lingvistika. Mežvuzovskij sbornik. Vypusk 2*, 82–89.

Zechner, K. (1996). Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In: *Proceedings of the International Conference on Computational Linguistics*. Copenhagen, 986–989.

Zubov, A. (2004). Formalization of the postup of singling out the basic text contents. *Journal of Quantitative Linguistics 11(1–2)*, 33–48.

Zubov, A. V., Čaplja, A. I., Čaplja, S. G. (1978). Avtomatičeskij otbor ključevych slov. In: *Strukturnaja i prikladnaja lingvistika. Vypusk I.* Leningrad: LGU, 198–205.

4.9 DÉLKA SLOVA A OBECNOST VÝZNAMU

Problém

Existuje souvislost mezi délkou a obecností významu slova?

Postup

Podle následujících pokynů si pomocí výkladového slovníku připravte přibližně 100 hyperonymních řetězců:

Hyperonymem základního lexému A je jiný obecnější lexém, který tvoří třídu, k níž A náleží. Například *nábytek* je hyperonymem *židle*, *budova* je hyperonymem *mrakodrapu*. Hyperonymum je obvykle součástí definice významu ve výkladovém slovníku. Uvažujte pouze substantiva, která vytvářejí hyperonymní řetězce. U některých jazyků uvádí hyperonymní řetězce Wordnet nebo podobné elektronické zdroje, u jiných si tyto řetězce musí badatel vytvořit sám. Doporučujeme tento základní postup.

(1) Neuvažujte žádné jiné vztahy než příslušnost k téže třídě, zejména neuvažujte meronymii (vztahy mezi částmi a celkem, např. *hlava = část těla*; *motor = část auta*, jelikož *tělo* není hyperonymem *hlavy* a zrovna tak *auto* není hyperonymem *motoru*).

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

(3) Snažte se vyvarovat cirkularity (ta se bohužel objevuje i ve Wordnetu).

(4) Vybírejte hyperonyma značně vysoké obecnosti či abstraktnosti, např. *celek*, *systém*, *bytost*, *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.

Stupeň obecnosti je možné odhadovat jako průměr úrovní, na nichž se dané substantivum vyskytlo. Např. v řetězci *kladivo – nástroj – nářadí – věc* by substantivum *nástroj* dosáhlo hodnoty obecnosti 2. Vyskytuje-li se slovo ve více než jednom řetězci, pak je možné jeho hodnoty obecnosti zprůměrovat.

Vypočtěte průměrnou obecnost (x) a délku (y) každého slova a hledejte mezi nimi nějakou zjevnou souvislost. Pokud vaše data vykazují značnou rozkolísanost, navyšte velikost výběrového souboru. Pokuste se návrhnout podobu případného trendu.

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 Linguistik. Ein internationales Handbuch. 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.

5 Textologie

5.1 BELZA-SKOROCHOĎKŮV KOEFICIENT ŘETĚZENÍ

Hypotéza

V odborných textech dochází k řetězení vět ve větší míře než v uměleckých textech. Ověřte tuto hypotézu.

Postup

Koeficient řetězení C_T vyjadřuje tendenci k vytváření plynulých sekvencí koherentních, tj. sémanticky souvisejících vět. Tento pojem je operacionalizován jako průměrná délka takovýchto řetězců vět v textu nebo souboru textů. Je tudíž definován jako

$$C_T = \frac{1}{S}\sum_{i=1}^{S} k_i \ ,$$

kde k_i představuje délku *i*tého řetězce (mohou existovat i řetězce o délce 1) a *s* je počet všech řetězců v textu.

Možnost sémantické spojitosti či nespojitosti po sobě jdoucích vět lze operacionalizovat různými způsoby, a to na základě koreference jako takové, anaforické koreference atd.

K ověření dané hypotézy si budete potřebovat definovat a operacionalizovat pojem sémantické spojitosti (můžete zkusit několik variant) a opatřit určitý počet pragmaticky homogenních textů, tj. textů téhož druhu, žánru, tematického zaměření apod. Poté u textů ve vašem souboru vypočtěte hodnoty C_T .

Belza (1971) uvádí, že ruské odborné texty vykazují hodnotu C_T = 7,4; populárně-vědné C_T = 6,6 a novinové či beletristické texty C_T = 5,3. Z toho plyne, že v ruských odborných textech je koherentní řetězec tvořen v průměru 7,4 větami.

Porovnejte texty podobného typu v různých jazycích a proveďte testy rovnosti. Koeficient C_T je prostým průměrem. K porovnání je proto možné využít t-test.

Další možné způsoby měření koherence lze najít v níže uvedené referenční literatuře.

Literatura

Bateman, J., Rondhuis, K. (1994). *Coherence relations: Analysis and specifications.* [odborná zpráva]. Darmstadt: GMD-IPSI.

Beaugrande, R.-A. de, Dressler, W. U. (1981). *Introduction to text linguistics*. London, New York: Longman.

Belza, M. I. (1971). K voprosu o nekotorych osobennostjach semantičeskoj struktury svjaznych tekstov. In: *Semantičeskie problemy avtomatizacii informacionnogo potoka*. Kiev, 58–73.

Dijk, T. A. van, Kintsch, W. (eds.) (1983). *Strategies of Discourse Comprehension*. New York, London: Academic Press.

Foltz, P. W. (1996). Comprehension, Coherence, and Strategies in Hypertext and Linear Text. In: Rouet, J.-F., Levonen, J. J., Dillon, A., Spiro, R. J. (eds.), *Hypertext and Cognition.* Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers, 109–136.

Fritz, G. (1982). *Kohärenz: Grundfragen der Dialoganalyse*. Tübingen: Narr.

Fritz, G. (1999). Coherence in Hypertext. In: Bublitz, W., Lenk, U., Eija, V. (eds.), *Coherence in Spoken and Written Discourse*. Amsterdam, Philadelphia: J. Benjamins, 221–232.

Haliday, M. A., Hassan, R. (1976). *Cohesion in English*. London: Longman.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science 3,* 67–90.

Hobbs, J. R. (1985). *On the coherence and structure of discourse*. [technical report]. Stanford, CA: Center for the Study of Language and Information, 85–37.

Rickheit, G., Schade, U. (2000). Kohärenz und Kohäsion. In: Brinker, K. Antos, G., Heinemann, W., Sager, S. F. (eds.), *Text- und Gesprächslinguistik – ein internationales Handbuch zeitgenössischer Forschung. 1. Halbband.* Berlin, New York: de Gruyter, 275–282.

Schade, U., Langer, H., Rutz, H., Sichelschmidt, L. (1991). Kohärenz als Prozeß. In: Rickheit, G. (ed.), *Kohärenzprozesse. Modellierung von Sprachverarbeitung in Texten und Diskursen*. Opladen: Westdeutscher Verlag, 7–58.

Schnotz, W. (1994). *Aufbau von Wissensstrukturen. Untersuchungen zur Kohärenzbildung bei Wissenserwerb mit Texten*. Weinheim: Beltz.

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

Strohner, H., Rickheit, G. (1990). Kognitive, kommunikative und sprachliche Zusammenhänge: Eine systemtheoretische Konzeption linguistischer Kohärenz. *Linguistische Berichte 125,* 3–23.

Stutterheim, C. von (1997). *Einige Prinzipien des Textaufbaus. Empirische Untersuchungen zur Produktion mündlicher Texte*. Tübingen: Niemeyer.

Wolf, F., Gibson, E. (2005). Representing discourse coherence: A corpus--based analysis. *Computational Linguistics 31(2),* 249–287.

Wolf, F., Gibson, E. (2005). *Coherence in natural language. Data structures and applications*. Cambridge: The MIT Press.

5.2 SHLUKOVÁNÍ AUTOSÉMANTIK

Problém

Prokažte, že u plnovýznamových slov dochází v rámci rankové distribuce slovních tvarů v textu k zvláštnímu shlukování v daných rankových intervalech.

Postup

Určete frekvence výskytu slovních tvarů (či lemmat) v textu a stanovte jejich rankovou distribuci. Vypočtěte h-bod a rozdělte text na intervaly o délce *h*. Následně spočítejte, kolik je v jednotlivých intervalech autosémantik. Výsledkem je monotónně rostoucí sekvence, kterou lze vyjádřit funkcí

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

(Popescu et al. 2009). V krátkých textech nemá tato sekvence příliš plynulý průběh. Nyní, když znáte *h* a *a*, definujte indikátor vyjadřující proporci autosémantik v daných intervalech (*autosemantic pace filling*, *APF*)

$$APF = a/h \quad (APF = autosemantic\ pace\ filling)$$

a vypočtěte jej pro několik textů. Tento index charakterizuje použití plnovýznamových slov v daném textu. Indikátor

$$AC = ak$$

pak současně vyjadřuje *autosémantickou kompaktnost* textu (Popescu et al. 2009).

Na základě těchto indikátorů lze zkoumat jazykový vývoj určitého autora, žánrů, stylů a také rozdílů mezi jazyky. Pro index *APF* si lze snadno vytvořit ověřovací test.

Pomocí výše uvedených vzorců dále rozvíjejte příslušnou teorii rozmístění a frekvence plnovýznamových slov v textech.

Literatura

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M. N. (2009). *Word frequency studies*. Berlin, New York: de Gruyter.

5.3 SÉMANTICKÁ REDUKCE V TEXTECH

Projekt

Určité procento slov v textu vždy tvoří slova polysémantická. Míru polysémie je možné zjistit pomocí dobrého slovníku, např. určením počtu jednotlivých významů označených arabskými či římskými číslicemi nebo písmeny. Nelze vyloučit ani další postupy, např. zkoumání významu určitého slova v každém z jeho kontextů v rámci nějakého megakorpusu.

Slova se v textu neobjevují izolovaně, ale jsou zanořena do svého specifického kotextu a kontextu, na jejichž základě jsou pak tato slova desambiguována, dochází k redukci a konkretizaci jejich sémantického potenciálu. Kotextový desambiguační efekt může vycházet z blízkých (afixy, kompozita, reduplikace) či vzdálenějších (frázových) prvků, na kontextovém efektu se pak podílí koreferenty a situace, k níž se odkazuje. Redukce polysémie způsobuje ve většině případů u určité sekvence slov její větší či menší monosémii, tj. hodnoty polysémie následných slov v textu obvykle tvoří sekvenci 1, 1, 1,… Postavíme-li takovou sekvenci do protikladu s polysémní sekvencí, jejíž hodnoty odpovídají slovníkové polysémii daných slov, zjistíme míru vzniklé redukce. Díky ustálené terminologii lze očekávat, že ve vědeckých textech dochází k menší redukci než například v poezii, kde se často pracuje se slovy, která mají podněcovat bohatou představivost. Text tedy může být prezentován jako sekvence sémantických redukcí.

Výsledné sekvence lze nazvat *sémantickými sekvencemi*. Takové sekvence mají množství obecných vlastností, ale lze je také různými způsoby rozčlenit na *P*-segmenty, které lze definovat jako neklesající sekvence čísel, např. 1, 1, 1, 2, 8. Jsou analogií jiných sekvencí tohoto druhu, např. sekvencí délek, frekvencí, polytextuality apod., které zavedli Köhler nebo Uhlířová (viz seznam referenční literatury). Tyto sekvence mají některé zajímavé vlastnosti: významově jednoznačným a vyčerpávajícím způsobem člení každý text, vyznačují se granularitou

mezi slovy a syntaktickými konstrukcemi a postihují syntagmatické textové struktury.

Z nastíněného stavu věcí vyplývají čtyři otázky, na něž je třeba hledat odpověď:

(1) Jaká je průměrná míra monosemizace v textu a jakou má monosemizace distribuci?

(2) Jaké jsou vlastnosti kompletní sekvence stupňů redukce v textu?

(3) Jaké jsou vlastnosti *P*-sekvencí, např. frekvence, délka, kombinace, distribuce?

(4) Jsme pomocí těchto vlastností schopni mechanicky rozlišovat žánry? Hledání odpovědí na všechny tyto otázky je zjevně úkolem pro celý tým badatelů, a proto o nich pojednáme jen v teoretické rovině.

Druhá skupina problémů vyplývá ze skutečnosti, že každé slovo ve slovníku náleží minimálně k jedné třídě slov, která s ním sdílí určité sémantické komponenty. V textu však někdy dochází ke ztrátě některých těchto komponentů (identifikátorů slovních tříd), protože slova se v textech užívají buď obecně, nebo mohou být specifikována prostřednictvím deixe, členů, kontextu, předložek apod. Jednotlivé jazyky využívají v tomto ohledu různé prostředky. Opět se zde objevuje stejný okruh problémů, tentokrát však jde o úbytek či nárůst počtu komponentů.

Prakticky všechny vlastnosti jazykových jednotek, které se dají kvantifikovat nebo alespoň nominálně klasifikovat, lze využít k vytváření symbolických nebo numerických sekvencí. Pokud je z definice možné tyto sekvence rozčlenit na kratší sekvence, označované jako segmenty, získáme tak abstraktní jazykové jednotky, které sice nebudou nositeli žádné formy ani významu, ale budou vyjádřením sekvence kvantifikovaných vlastností. Jelikož je tímto způsobem možné postihnout všechny vlastnosti jednotek, nabízí se tak možnost vzniku nové disciplíny.

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: Slovak Academic Press, 145–152.

Köhler, R. (2008). Word length in text. A study in the syntagmatic dimension. In: Mislovičová, S. (ed.), *Jazyk a jazykoveda v pohybe.* Bratislava: Veda, 416–421.

Köhler, R. (2008). Sequences of linguistic quantities. Report on a new unit of investigation. *Glottotheory 1(1)*, 115–119.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, B., Schmidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications.* Berlin, Heidelberg: Springer, 637–646.

Uhlířová, L. (2009). Word frequency and position in sentence. *Glottometrics 14*, 1–21.

5.4 RANKOVÁ DISTRIBUCE A DÉLKA KŘIVKY

Problém

Prokažte, že délka křivek rankových distribucí koreluje s entropií.

Postup

Vypočtěte rankovou distribuci slovních tvarů v několika textech. Vypočtěte relativní frekvence a použijte je k výpočtu Shannonovy entropie

$$H = -\sum_{x=1}^{V} p_x \log_2 p_x \ ,$$

kde V je počet různých slovních tvarů a p_x jsou relativní frekvence. Následně vypočtěte délku křivky distribuce pomocí vzorce

$$L = \sum_{x=1}^{V-1} [(f_x - f_{x+1})^2 + 1]^{1/2} \ ,$$

kde f_x jsou absolutní frekvence. Demonstrujte, že mezi H a L existuje přinejmenším nějaká korelace, a určete formu takového vzájemného vztahu.

Literatura
žádná

5.5 BOHATOST SLOVNÍ ZÁSOBY DLE POPESCA

Problém
Vyhodnoťte frekvence slov v různých textech a pomocí Popescových indikátorů vypočtěte bohatost slovní zásoby (Popescu et al. 2009).

Postup
Bohatost slovní zásoby lze vyhodnocovat mnoha různými způsoby. Nejprve si prostudujte dostupnou literaturu a zpracujte si přehled jednotlivých indikátorů a postupů. Posuďte jejich výhody a nevýhody. Následně použijte všechny indikátory, které uvádějí Popescu et al. (2009), a své výsledky porovnejte s výsledky uváděnými v dané publikaci.

Proveďte testy rovnosti, rozdělte vybrané texty do různých skupin podle jejich bohatosti a interpretujte výsledky, k nimž jste dospěli.

Literatura
Brunet, É. (1978). *Le vocabulaire de Jean Giraudoux. Structure et evolution*. Genève: Slatkine.

Cossette, A. (1994). *La richesse lexicale et sa mesure*. Paris: Champion.

Dugast, D. (1980). La mesure de la richesse lexicale: une esquisse historique. *Verbum 3(1),* 115–134.

Honore, T. (1979). Some simple measures of richness of vocabulary. *ALLC Journal 7,* 172–177.

Kuraszkiewicz, W. (1963). *La richesse du vacabulaire dans quelques grands texts polonaise en vers.* Wroclaw: Ossolineum.

Ménard, N. (1983). *Mesure de la richesse lexicale*. Paris: Slatkine.

Ménard, N., Santerre, L. (1979). La richesse lexicale individuelle comme marquer sociolinguistique. *Cahiers der linguistique 1,* 165–188.

Muller, Ch. (1968). Mesure de la richesse lexicale. *Travaux de linguistique et de littérature 6,* 73–84.

Muller, Ch. (1971). Sur la mesure de la richesse lexicale. Théorie et experience, homage a René Michéa. *Études de linguistique appliqué,* 74–87.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M. N. (2009). *Word frequency studies.* Berlin, New York: de Gruyter.

Ratkowsky, D. A., Hantrais, L. (1975). Tables for comparing the richness and structure of vocabulary in texts of different lengths. *Computers and Humanities 9,* 69–75.

Ratkowsky, D. A., Halstead, M. H., Hantrais, L. (1980). Measuring vocabulary richness in literary works. A new proposal and re-assessment of some earlier measures. *Glottometrika 2,* 125–147.

Tešitelová, M. (1972). On the so-called vocabulary richness. *Prague Studies in Mathematical Linguistics 3,* 103–120.

Thoiron, P. (1986). Indice de diversité et mesure de la richesse lexicale. In: Muller, Ch., *Méthodes quantitatives et informatiques dans l´étude des texts. Colloque international de CNRS a l´Université de Nice, 5–8 juin 1985.* [Festschrift]. Paris: Champion, 637–646.

Thoiron, P., Labbé, D., Serant, D. (eds.) (1988). *Études sur la richesse et la structure lexicale*. Paris, Geneve: Champion, Slatkine.

Wimmer, G., Altmann, G. (1999). Review article: On vocabulary richness. *Journal of Quantitative Linguistics 6(1),* 1–9.

5.6 ALITERACE

Hypotéza

Každá báseň vykazuje určitou míru nenáhodné aliterace.

Postup

Aliterace je opakování shodných fonémů (hlásek, písmen, atd.) na začátku úvodních slov verše. K ověření výše formulované hypotézy musíte nejprve určit relativní frekvence fonémů (hlásek, písmen, atd.) ve zkoumaném jazyce. Pokud to případně nebude možné, budete muset použít frekvence fonémů z dané básně. Nechť relativní frekvence fonému *i* je p_i, nechť počet slov ve verši je *n* a počet slov začínajících fonémem *i* je *r*. Pravděpodobnost nalezení přesně *r* slov začínajících na *i* je pak dána vzorcem

$$(1) \qquad P(X_i = r) = \binom{n}{r} p_i^r q_i^{n-r}$$

a pravděpodobnost, že *r* nebo více slov začíná na *i*, je dána

$$(2) \qquad P(X_i \geq r) = \sum_{x=r}^{n} \binom{n}{x} p_i^x q_i^{n-x} \ ,$$

kde $q_i = 1 - p_i$. Pokud je *P* (ve vzorci 2) menší než 0,05; pak můžete přijmout hypotézu, že u daného fonému v daném verši dochází k aliteraci. U všech veršů proveďte test na všechny fonémy, které se v jednotlivých verších vyskytují minimálně dvakrát. Stanovte aliterační index. Proveďte srovnání jednotlivých básní, autorů, lyrických a epických básní a pokuste se vysledovat určitý diachronní vývoj.

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.

5.7 ALITERAČNÍ STRUKTURA

Hypotéza

Každá báseň má nějakou nenáhodnou aliterační strukturu. Ověřte tuto hypotézu.

Postup

Pokud se na začátku úvodních slov verše opakuje více než jeden foném, např. foném *i* se vyskytuje na začátku dvou slov a foném *j* na začátku tří slov, pak se postup uvedený v části 5.6, „Aliterace", poněkud komplikuje. Musíme zde obecně pracovat s multinomickým rozdělením, přičemž v případě dvou opakujících se fonémů použijeme trinomické rozdělení. Nechť p_i je pravděpodobnost fonému *i*, p_j pravděpodobnost fonému *j* a $1\text{-}p_i\text{-}p_j$ pravděpodobnost ostatních fonémů. Pak pravděpodobnost, že ve verši s *n* slovy se vyskytuje přesně k_i slov začínajících fonémem *i*, k_j slov začínajících fonémem *j* a $n\text{-}k_i\text{-}k_j$ slov začínajících jakýmkoli jiným (neopakujícím se) fonémem, je dána vzorcem

$$(1) \quad P(X_i = k_i, X_j = k_j, X_{n-k_i-k_j} = n - k_i - k_j) = \\ = \frac{n!}{k_i! k_j! (n - k_i - k_j)!} p_i^{k_i} p_j^{k_j} (1 - p_i - p_j)^{n-k_i-k_j} .$$

Pro výpočet dané a krajnější pravděpodobnosti sečteme jednotlivé pravděpodobnosti:

$$(2) \quad P(X_i \geq k_i, X_j \geq k_j, X_{n-k_i-k_j} = n - k_i - k_j) = \\ = \sum_{\substack{x_i \geq k_i \\ x_j \geq k_j}} = \frac{n!}{x_i! x_j! (n - x_i - x_j)!} p_i^{x_i} p_j^{x_j} (1 - p_i - p_j)^{n - x_i - x_j} .$$

Jakmile budete mít vypočítánu míru aliterace pro jednotlivé verše, zaměřte se na průběh aliterace v básni a pomocí vhodného indikátoru vyjádřete míru této aliterace. Vzorec (2) lze také použít k výpočtu míry aliterace na začátku veršů.

Literatura

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

5.8 DISORTATIVITA AUTOSÉMANTIK

Hypotéza

Asociační graf autosémantických slov v textu je disortativní. Ověřte tuto hypotézu.

Postup

Nejprve nahraďte všechna zájmena v textu slovy, která zastupují. Poté z textu eliminujte všechna pomocná slova (auxiliáry), ponechejte pouze autosémantika. Pomocí metody představené v knize *Kvantitativní lingvistika. Vybrané problémy 1* (Strauss et al. 2014, kap. „Asociační graf textu“) vypočtěte koincidenci autosémantik v rámci věty. Pro každý vrchol (plnovýznamové slovo) určete počet jeho hran (= počet spojení daného slova s jinými autosémantiky v grafu). Určete, zda slova s vyšším počtem hran vykazují větší propojenost spíše se slovy s vyšším počtem hran než se slovy s nižším počtem hran. V prvním případě je

graf asortativní, ve druhém disortativní. Test provedete jednoduše vypočtením korelace počtů hran u jednotlivých vrcholů.

Prokažte, že texty speciálního zaměření (např. vědecké) jsou asortativnější než poetické texty (a naopak). Pokuste se určit míru asortativity u různých druhů textu a zkoumejte tento problém z diachronního hlediska.

Literatura

Newman, M. E. J. (2001). *Clustering and preferential attachment in growing networks*. [Dostupné z: arXiv:cond-mat/0104209v1 (cit. 11. dubna 2011).]

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

Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review E 67*, article: 026126.

Newman, M. E. J., Park, J. (2003). Why social networks are different from other types of networks. P*hysical Review E* 68, article: 036122.

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

5.9 SUPERHREB

Hypotéza

Mezi rovinami textu se nacházejí super-jednotky, jež jsou tvořeny hreby.

Postup

Uvažujeme-li *hreby* (syntaktické konstrukty) jako agregáty vět obsahujících totéž slovo nebo tentýž symbol nebo tutéž významovou jednotku (včetně synonym) atd., pak podle Menzerathova zákona může existovat rovina tvořená *„superhreby"*, tj. jednotkami, jejichž součástmi jsou *„hreby"*. Vytvořte jednotky tohoto druhu a nalezněte příslušnou rovinu.

Literatura

Hřebíček, L. (1992). *Text in communication: supra-sentence structures*. Bochum: Brockmeyer.

Schwarz, C. (1996). The distribution of aggregates in texts. *ZET–Zeitschirft für Empirische Textforschung 2,* 62–66.

5.10 ZLATÝ ŘEZ (1)

Hypotéza

Radiány úhlu rankové distribuce, které vyjadřují tzv. „autorovo hledisko", nejsou nikdy menší než hodnota zlatého řezu 1,618.

Postup

Vypočtěte rankovou distribuci slovních tvarů v textu. Nechť *r* = rank, *f(r)* = frekvence při ranku *r*. Vypočtěte Hirsch-Popescův h-bod podle vzorce

$$h = \begin{cases} r, & \textit{pokud existuje případ, kdy } r = f(r) \\ \dfrac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)}, & \textit{pokud neexistuje případ, kdy } r = f(r) \end{cases}$$

tj. $h = r$, pokud existuje rank *r*, jehož frekvence *f(r)* se rovná tomuto ranku, pokud ne, vezměte dva sousední ranky r_i and r_j, aby platilo, že $r_i < f(r_i)$ a $r_j = r_i + 1 > f(r_j)$, a vyřešte druhou část vzorce. Pomocí přímky spojte tento bod $P(h, h)$ s body $P[1,f(1)]$, tj. s nejvyšší frekvencí, a $P(V, 1)$, tj. s frekvencí při nejvyšším ranku (V = slovní zásoba či inventář). Úhel α související s *h*-bodem se označuje jako „autorovo hledisko" (srov. Popescu, Altmann 2007). Vypočtěte nejprve cos α daný jako

$$\cos\alpha = \frac{-\left[(h-1)(f(1)-h) + (h-1)(V-h)\right]}{\left[(h-1)^2 + (f(1)-h)^2\right]^{1/2}\left[(h-1)^2 + (V-h)^2\right]^{1/2}}$$

poté α a na závěr radiány

$$\alpha \text{ rad} = 2\pi\alpha/360.$$

Úhel α rad se musí rovnat minimálně π/2 = 1,57; ale vždy je větší než 1,618; což znamená, že pro texty je jeho dolní mezí zlatý řez.

Proveďte tuto analýzu na větším množství textů a ověřte tuto hypotézu. Současně věnujte pozornost maximální hodnotě α rad. Teoreticky se rovná π, avšak z empirického hlediska jeho hodnota není známa.

Literatura

Popescu, I.-I., Altmann, G. (2007). Writer´s view of text generation. *Glottometrics 15*, 71–81.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M. N. (2009). *Word frequency studies*. Berlin, New York: de Gruyter.

5.11 PODIVNÝ ATRAKTOR AUTOROVA HLEDISKA

Hypotéza

Radiány *α* „autorova hlediska" leží v podivném atraktoru.

Postup

Vyřešte nejprve problém dle zadání v části 5.10, „Zlatý řez", a shromážděte výsledky z většího množství textů. Zohledněte rovněž délku textů *N*. Do kartézské souřadnicové soustavy zadejte body <*N*, radián *α*>. Dostaneme plochu podobnou bumerangu, pravděpodobně s několika odlehlými body. Nalezněte minimálně nějakou funkci vystihující průběh jednotlivých bodů. Pokud to bude možné, pomocí soustavy dvou diferenciálních rovnic stanovte přibližnou oblast polohy bodů.

Literatura

Popescu, I.-I., Altmann, G. (2007). Writer´s view of text generation. *Glottometrics 15*, 71–81.

5.12 ARISTOTELOVY KATEGORIE

Problém

Proveďte analýzu textu z hlediska Aristotelových kategorií.

Postup

Uvažujte tyto aristotelské kategorie:

podstatu	– co je něco? (lavice, dívka)
kvantitu	– jak je něco velké? (dva metry)
kvalitu	– jaké je něco? (schopné, zelené)
vztah	– v jakém vztahu je to k něčemu? (větší)
místo	– kde je to? (ve škole)
čas	– kdy je to? (dnes, zítra)
polohu	– v jaké je to pozici? (sedí to, visí to)
vlastnictví/habitus	– co to má? (je to ozbrojené, má to klobouk)
činnost	– co to dělá? (běží to, řeže to)
trpnost (pasivně přijímáno)	– co se tomu děje? (řežou to, pálí to)

„Hodnota těchto kategorií je dnes běžně vnímána jako čistě historická, zčásti z důvodu všeobecně odmítaného Aristetolova pojetí podstaty. Toto odmítání často pramení z nepochopení pravého významu tohoto pojetí, a sice že podstata je to, co existuje samo o sobě a nikoli v něčem jiném.“ [Dostupné z: http://en.wikipedia.org/wiki/Category_(philosophy) (cit. 1. prosince 2008)].

Prostřednictvím těchto kategorií je možné text rozložit nikoli na slova, ale mnohdy na syntagmata, fráze atd. Určete počet jednotlivých kategorií a demonstrujte jejich prostřednictvím rozdíly mezi texty. Nebude-li počet kategorií stačit, zaveďte další.

Literatura

Aristoteles. (1953). *Metaphysics*. Přel. Ross, W. D., Oxford University Press.

Aristoteles. (2004). *Categories*. Přel. Edghill, E. M., University of Adelaide Library.

5.13 SKINNERŮV EFEKT

Problém

Pokud je výskyt jazykových jednotek v textu autostimulován v tom smyslu, že s výskytem určité jednotky roste současně pravděpodobnost jejího výskytu v těsné blízkosti téže jednotky (Skinner 1939, 1941, 1957), pak z toho automaticky vyplývá následující hypotéza: verše básně nacházející se ve vzájemné blízkosti jsou foneticky podobnější než verše vzdálenější. Ověřte tuto hypotézu.

Postup

Vyberte si nějakou delší báseň v libovolném jazyce a foneticky ji přepište řádek po řádku, případně proveďte její morfematickou transkripci. Definujte měřítko fonetické podobnosti mezi verši, vypočtěte průměrné hodnoty podobnosti pro vzdálenosti 1, 2, 3, … a pokuste se zjistit, zda průběh hodnot podobnosti klesá s rostoucí vzdáleností.

Aplikujte tuto analýzu jak na umělecké texty, tak na lidovou poezii.

Je možné učinit závěr, že text vyznačující se touto pravidelností vznikl spontánněji než text, který tuto pravidelnost nevykazuje?

Pokuste se tento jev sledovat diachronně. Je aliterace důsledkem tohoto jevu?

Literatura

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

Bunde, A., Eichner, J. F., Kantelhardt, J. W., Havlin, S. (2005). Long-term memory: a natural mechanism for the clustering of extreme events and anomalous residual times in climate records. *Physical Review Letters 94*, article: 048701.

Corral, A., Ferrer-i-Cancho, R., Diaz-Guilera, A. (2009). *Universal complex structures in written language*. [Dostupné z: http://arxic.org/abs/0901.2924 (cit. 7. ledna 2009).]

Skinner, B. F. (1939). The alliteration in Shakespeare´s sonnets: A study in literary behavior. *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.

Skinner, B. F. (1957). *Verbal Behaviour*. Acton: Copley Publishing Group.

5.14 SCHÉMA <I,J>

Problém

Zjistěte rankovou distribuci slovních tvarů z většího množství textů. V rámci kartézské soustavy souřadnic <*I,J*> vyznačte indikátory *I* a *J*. Zaměřte se na pozice textů různých žánrů a také na jejich pozici v historickém sledu. Popište, co jste vypozorovali.

Postup

Problematika pozic indikátorů *I* a *S* (Ord 1972) u různých distribučních dat je v lingvistice dobře známa (srov. Strauss et al. 2014, kap. „Ordovo kritérium"). Popescu, Mačutek, Altmann (2009) představili nový přístup, v němž se využívá entropie distribuce.

Definujte

$$I = \frac{m_2}{m_1'} = \frac{s^2}{\overline{x}} \ ,$$

které je identické s Ordovým *I*. Jelikož je entropie v rankových distribucích zároveň indikátorem šikmosti, souřadnice *J* je definována jako

$$J = \frac{H}{s_{\overline{x}}} \ ,$$

tj. při entropii definované jako

$$H = -\sum_{i=1}^{V} p_i \log_2 p_i \ ,$$

kde V je velikost slovní zásoby (a nejvyšší rank), p_i jsou relativní frekvence a s_x je směrodatná odchylka od průměru.

(1) Je pomocí tohoto schématu možné rozlišovat mezi výrazně analytickými a výrazně syntetickými jazyky?

(2) Je možné zřetelně odlišit anglické texty od textů v němčině nebo slovanských jazycích?

(3) Lze pozorovat nějaký vývoj v díle jednoho konkrétního autora?

(4) Je možné na datech z historického korpusu demonstrovat vývoj němčiny k analytismu?

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., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M. N. (2009). *Word frequency studies*. Berlin, New York: de Gruyter.

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). *Aspects of word frequencies*. Lüdenscheid: RAM.

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

5.15 TEXTOVÁ KOHEZE (1)

Problém

Určete blokovou distribuci anafory a katafory v textech různého druhu.

Postup

V analogii k blokové distribuci funkčních slov (popisované Frumkinovou [1962] a dalšími – viz literatura) a syntaktických konstrukcí/funkcí (srov. Köhler 2001) určete počet textových bloků (zkuste bloky o velikosti 10, 30, 50 a 100 slov) s výskytem 0, 1, 2, … anafor a katafor. Aplikujte negativně hypergeometrické rozdělení („Frumkinové zákon“) na daná data. Pozorujte závislost hodnot parametrů na délce bloků, počtu bloků a typu kategorie.

Negativně hypergeometrické rozdělení je definováno jako

$$P_x = \frac{\binom{M+x-1}{x}\binom{N-M+n-x-1}{n-x}}{\binom{K+n-1}{n}}, \; x = 0, 1, \ldots, n$$

kde K, M a n jsou parametry.

Zjistěte, zda lze tyto parametry považovat za charakteristiky textu.

Za jakých specifických okolností by bylo možné uplatnit limitní případy negativně hypergeometrického rozdělení, konkrétně Poissonova, binomického a negativně binomického rozdělení?

Pokud se některé parametry jejich funkcí jeví jako vhodné k vyjádření textové koheze, vytvořte příslušný indikátor a demonstrujte jeho vlastnosti.

Literatura

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

Altmann, G., Burdinski, V. (1982). Towards a law of word repetitions in text-blocks. *Glottometrika 4,* 146–167.

Bektaev, K. B., Luk´janenkov, K. F. (1971). O zakonach raspredelenija edinic pis´mennoj reči. In: Piotrowski, R. H. (ed.), *Statistika reči i avtomatičeskij analiz teksta*. Leningrad: Nauka, 47–112.

Brainerd, B. (1972). Article use as an indirect indicator of style among English-language authors. In: Jäger, S. (ed.), *Linguistik und Statistik*. Braunschweig: Vieweg, 11–32.

Francis, I. S. (1966). An exposition of a statistical approach to Federalist dispute. In: Leed, J. (ed.), *The computer and literary style*. Kent, Ohio: Kent State University Press, 38–78.

Frumkina, R. M. (1962). O zakonach raspredelenija slov i klassov slov. In: Mološnaja, T. N. (ed.), *Strukturno-tipologičekie issledovanija*. Moskva: Akademija Nauk SSSR, 124–133.

Köhler, R. (2001). The distribution of some syntactic construction types in text blocks. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček*. Trier: Wissenschaftlicher Verlag, 136–148.

Maškina, L. E. (1968). *O statističeskich metodach issledovanija leksiko-gramatičeskoj distribucii*. Minsk, Diss.

Mosteller, F., Wallace, D. L. (1964). *Inference and disputed authorship*: *The Federalist*. Reading, Mass: Addison-Wesley.

Paškovskij, V. E., Srebrjanskaja, I. I. (1971). Statističeskie ocenki pis´mennoj reči bol'nych šizofreniej. In: *Inženernaja lingvistika*. Leningrad.

Piotrowski, R. G. (1984). *Text, Computer, Mensch*. Bochum: Brockmeyer.

Wimmer, G., Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

5.16 TEXTOVÁ KOHEZE (2)

Hypotéza

Vzdálenosti mezi anaforami a kataforami představují monotónně sestupnou sekvenci.

Postup

Určete vzdálenosti mezi všemi anaforami (kataforami) v textu pomocí číselného údaje označujícího počet slov, kterými jsou odděleny, a stanovte frekvenční distribuci těchto vzdáleností (tj. počtu výskytů vzdáleností o velikosti 0, 1, 2, … slov). Proveďte aproximaci této sekvence za použití Zipf-Aleksejevovy funkce

$$y = ax^{-b-c\ln x},$$

kde x je vzdálenost ($x = 0, 1, 2, \ldots$ nebo $x = 1, 2, 3, \ldots$ v závislosti na použité definici vzdálenosti), y je počet výskytů této vzdálenosti a a, b, c jsou parametry. V případě nevhodnosti výše uvedené funkce nalezněte funkci vhodnější. (Nezapomeňte, že $x^0 = 1$)

Existuje nějaká spojitost mezi funkcí vzdálenosti (např. jejím průměrem) a parametry negativně hypergeometrického rozdělení, o nichž byla řeč v části 5.14, „Schéma <I,J>"? Pokud ano, určete, o jaký typ provázanosti se jedná, a formálně jej vyjádřete.

Literatura

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.

Skinner, B. F. (1957). *Verbal behaviour.* Acton: Doplet.

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

5.17 TEXTOVÁ KOHEZE (3)

Hypotéza

Frekvenční distribuce gramatických funkcí anafory a katafory se řídí rozdělením pravděpodobnosti náležejícím do třídy diversifikačních rozdělení.

Postup

Určete gramatickou funkci každé anafory (katafory) v textu a počet výskytů těchto jednotlivých funkcí. Aplikujte na tato data příslušné rozdělení pravděpodobnosti. Zdůvodněte toto rozdělení.

Literatura

Alekseev, P. M. (1978). O nelinejnych formulirovkach zakona Cipfa. In: Piotrovskij, R. G. (ed.), *Statistika reči avtomatičeskij analiz teksta. Moskva-Leningrad: Naučnyj sovet po kompleksnoj probleme „Kibernetika"*. AN SSSR, 53–65.

Altmann, G. (1985a). Semantische Diversifikation. *Folia Linguistica 19,* 177–200.

Altmann, G. (1985b). Die Entstehung diatopischer Varianten. Ein stochastisches Modell. *Zeitschrift für Sprachwissenschaft 4,* 139–155.

Altmann, G. (1991). Modelling diversification phenomena in language. In: Rothe, U. (ed.), *Diversification processes in language: Grammar.* Hagen: Rottmann, 33–46.

Altmann, G. (1996). Diversification processes of the word. *Glottometrika 15,* 102–111.

Altmann, G., Best, K.-H., Kind, B. (1987). Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. *Glottometrika 8,* 130–139.

Beöthy, E., Altmann, G. (1984a). The diversification of meaning of Hungarian verbal prefixes. II. „ki-". *Finnisch-Ugrische Mitteilungen 8,* 29–37.

Beöthy, E., Altmann, G. (1984b). Semantic diversification of Hungarian verbal prefixes. III. „föl-", „el-", „be-". *Glottometrika 7,* 45–56.

Best, K.-H. (1994). Word class frequency in contemporary German short prose texts. *Journal of Quantitative Linguistics 1,* 144–147.

Best, K.-H. (2009). Diversifikation des Phonems /r/ im deutschen. *Glottometrics 18,* 26–31.

Haight, F. A. (1966). Some statistical problems in connection with word association data. *Journal of Mathematical Psychology 3,* 217–233.

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik. Aufbau eines lexikalischen Basismodells*. Trier: WVT.

Hoffmann, L. (2000). Anapher im Text. In: Brinker, K., Antos, G., Heinemann, W. (eds.), *Text- und Gesprächslinguistik. Linguistics of text and conversation.* Berlin, New York: de Gruyter, 295–304.

Horvath, W. J. (1963). A stochastic model for word association tests. *Psychological Review 70,* 361–364.

Hřebíček, L. (1996). Word associations and text. *Glottometrika 15,* 96–101.

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

Köhler, R. (1987). Systems theoretical linguistics. *Theoretical Linguistics 14,* 241–257.

Köhler, R. (1989). Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation. *Glottometrika 11,* 1–18.

Köhler, R. (1990). Elemente der synergetischen Linguistik. *Glottometrika 12,* 179–17.

Rothe, U. (ed.) (1991). *Diversification processes in language: Grammar.* Hagen: Rottmann, 47–55.

Laufer, J., Nemcová, E. (2009). Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18,* 13–25.

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18,* 32–39.

Rothe, U. (ed.) (1991). *Diversification processes in language: Grammar.* Hagen: Rottmann, 85–91.

Rothe, U. (1986). *Die Semantik des textuellen et.* Frankfurt: Lang.

Rothe, U. (ed.) (1991). *Diversification processes in language: Grammar.* Hagen: Rottmann.

Ziegler, A. (1998). Word class frequencies in Brasilian-Portuguese press texts. *Journal of Quantitative Linguistics 4,* 269–280.

5.18 HAPAX LEGOMENA A MARKOVOVY ŘETĚZCE

Hypotéza

Vzdálenost mezi hapax legomeny v textu je Markovovým řetězcem prvního řádu (Popescu et al. 2009: 227). Ověřte tuto hypotézu.

Postup

Vypočtěte frekvence slov v textu. Poté nahraďte hapax legomena nějakým symbolem, např. 1, a ostatní slova 0. Vzdálenost mezi dvěma hapax legomena je počet ostatních slov, která leží mezi nimi (= počet nul mezi dvěma jedničkami). Je-li tato sekvence Markovovým řetězcem, pak distribuce vzdáleností (Y) odpovídá modifikovanému geometrickému rozdělení

$$P(Y = k) = \begin{cases} 1 - \alpha, & \text{při } k = 0 \\ \alpha p q^{k-1}, & \text{při } k = 1, 2, 3 \end{cases}.$$

Aplikujte danou distribuci na frekvence vzdáleností. Pomocí softwaru Altmann-Fitter lze toto provádět iterativně, o bodových odhadech pojednávají blíže Strauss et al. (1984). Porovnejte parametry a a p ($q = 1 - p$) v různých textech a prostřednictvím intervalů mezi parametry charakterizujte jednotlivé texty a žánry.

Proveďte analýzu textů v různých jazycích a uveďte, zda daná hypotéza platí. Pokud ano, uveďte dané parametry do souvislosti s některou jinou vlastností daného jazyka, např. syntetismem/analytismem.

Pokud lze hypotézu odmítnout, nepřecházejte k vyššímu řádu Markovova řetězce, protože tento postup by vedl pouze k dalším modifikacím geometrického rozdělení, hledejte jiné řešení (srov. část 5.19, „Frekvenční sekvence slov").

Srov. také Strauss et al. (2014, kap. „Vzdálenosti mezi stejně dlouhými větami").

Literatura

Brainerd, B. (1976). On the Markov nature of the texts. *Linguistics 76*, 5–30.

Popescu, I.- I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M. N. (2009). *Word frequency studies*. Berlin, New York: de Gruyter.

Strauss, U., Sappok, Ch., Diller, H. J., Altmann, G. (1984). Zur Theorie der Klumpung von Textentitäten. *Glottometrika 7*, 73–100.

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

5.19 FREKVENČNÍ SEKVENCE SLOV

Problém

Sekvence frekvencí slov vykazuje určitou pravidelnost. Identifikujte tuto pravidelnost a využijte ji k jejich porovnání.

Postup

Vypočtěte frekvence slov v textu (lemmat nebo slovních tvarů). Pro tento účel můžete využít některý z volně dostupných programů na počítání slov. Poté nahraďte slova v textu jejich frekvencemi. Interpunkci zanedbávejte. Výsledkem je určitá časová řada tvořená sekvencí frekvencí.

(a) Pomocí Fourierovy analýzy vyjádřete oscilaci těchto frekvencí.

(b) Vypočtěte Hurstův exponent pro tuto řadu, srov. Strauss et al. (2014, kap. „Hurstův exponent").

(c) Proveďte srovnání textů na základě výsledků z (a) a (b).

(d) Stanovte distribuci rozdílů mezi sousedními frekvencemi a najděte diskrétní teoretické rozdělení. Danou distribuci zdůvodněte pomocí gramatických, typologických, psychologických a dalších argumentů.

(e) Vypočtěte některé vlastnosti empirického rozdělení a porovnejte je s jinými texty.

(f) Z distribuce v (c) vyvoďte některé typologické závěry.

Literatura

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., 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.

Spiegel, M. R. (1974). *Fourier Analysis*. New York: McGraw-Hill.

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

5.20 ZLATÝ ŘEZ (2)

Problém

Za využití Popescovy-Altmannovy (2009) metody prokažte asymptotickou existenci zlatého řezu v textech.

Postup

U textů různé délky, až v rozsahu celých románů, vypočtěte rankovou distribuci lemmat nebo slovních tvarů a současně vypočtěte pro každý text h-bod (viz část 5.10, „Zlatý řez [1]“), délku křivky L (viz část 6.1, „Délka křivky a typologie“) a maximální délku křivky podle

$$L_{max} = V - 1 + f(1) - 1 \ ,$$

kde V je velikost slovní zásoby (tj. nejvyšší rank) a $f(1)$ je nejvyšší frekvence. Na základě těchto veličin vypočtěte dva textové indikátory p a q definované jako

$$p = \frac{L_{max} - L}{h - 1} \qquad \text{a} \qquad q = \frac{L_{max} - L}{\sqrt{N}} \ .$$

Sečtěte je, tj. vypočítejte

$$p + q = (L_{max} - L)\left(\frac{1}{\sqrt{N}} + \frac{1}{h-1}\right)$$

a prokažte souběh výsledné hodnoty se zlatým řezem 1,618… Zaměřte se blíže na chování *p* i *q* a věnujte pozornost způsobu konvergence *p+q*. Bude-li to možné, pokuste se tento zvláštní jev zdůvodnit lingvisticky, formulujte jeho matematická východiska a identifikujte jednoduché funkce nebo intervaly pro *p*, *q* a *p+q* ve vztahu ke zkoumaným datům. Své výsledky porovnejte se závěry, ke kterým dospěli Tuzzi et al. (2009).

Následně proveďte analýzu rankové sekvence jiné jednotky vyskytující se napříč celým textem, např. slovních druhů, vypočtěte výše zmíněné indikátory a zjistěte, zda i v tomto případě dochází k souběhu *p+q* se zlatým řezem.

Na závěr se zaměřte na jednotlivé slovní druhy. Rozdělte je na vhodné dílčí kategorie, v případě zájmen by takovou skupinu dílčích kategorií představovaly např. zájmena osobní, ukazovací, vztažná, tázací apod. Stanovte rankovou sekvenci těchto dílčích kategorií a proveďte analýzu změny *p*, *q*, *p+q* na této úrovni. Je možné pozorovat nějaký trend nebo určitou soběpodobnost apod.?

Literatura

Popescu, I.-I., Altmann, G. (2009). A modified text indicator. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Problems of quantitative text analysis*. Černivci: ČNU, 13–39.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). The golden section in texts. ETC – *Empirical Text and Culture Research 4*, 30–41.

6 Typologie a univerzálie

6.1 DÉLKA KŘIVKY A TYPOLOGIE

Problém

Je vztah mezi délkou křivky rankové distribuce slovních tvarů a největší frekvencí typologickým indikátorem? Zaměřte se podrobněji na tento problém.

Postup

Zvolte si několik textů a sestavte rankové sekvence prvních 50 slovních tvarů z každého z nich. Určete jednotlivé $f(1)$, tj. frekvence slov s rankem 1, a vypočtěte délky křivek podle

$$L = \sum_{r=1}^{V-1} \left[(f(r) - f(r+1))^2 + 1 \right]^{1/2},$$

kde $f(r)$ je frekvence při ranku r a V je slovní zásoba (= počet typů slovních tvarů). Zaznamenejte si L a $f(1)$. Poté tentýž postup opakujte, ale tentokrát s prvními 100 slovními tvary. Opět si zaznamenejte $f(1)$ a L. Pokračujte maximálně do 1 000 slov, měli byste nakonec získat 20 bodů $< L_i, f(1)_i >$. Jakmile budete mít tyto veličiny pro několik textů, vypočtěte funkci

$$L = a f_1^b \ .$$

Tentýž postup aplikujte na texty v několika jazycích. Čím je jazyk syntetičtější, tím je funkce strmější, tj. tím je větší parametr b. Zaměřte pozornost na chování parametru b a vysvětlete jeho roli z typologického hlediska. Proveďte analýzu textů v příbuzných jazycích, např. románských.

Literatura

Popescu, I.-I., Mačutek, J., Altmann, G. (2008). Word frequency and arc length. *Glottometrics 17,* 18–44.

6.2 DÉLKA MORFŮ

Hypotéza

Délka morfů vykazuje pravidelnou distribuci (Saporta 1966, Best 2001). Ověřte tuto hypotézu.

Postup

Saporta si všiml, že délka španělských morfů měřená z hlediska počtu fonémů vykazuje velmi pravidelné schéma a položil si otázku ohledně možné univerzálnosti tohoto jevu. Dospěl k následujícím datům:

TABULKA 6.2.1, Délka morfů měřená z hlediska počtu fonémů

Počet fonémů	Počet různých morfů
0	6
1	59
2	97
3	307
4	387
5	327
6	261
7	143
8	64
9	19
10	4
11	1

Počet fonémů	Počet různých morfů
12	2
13	1
14	1

Najděte diskrétní rozdělení vyjadřující frekvence morfů jednotlivých délek a pokuste se zjistit „… jaký jiný faktor než náhoda by se mohl podílet na vzniku takového rozdělení". (Saporta 1969: 69).

Byla-li aplikace rozdělení na daná data úspěšná, z nějakého jiného jazyka nebo skupiny jazyků vytvořte náhodný výběrový soubor čítající cca 1 500 morfů a výše uvedené výsledky zobecněte. Z formálního a metodologického hlediska se tento problém nijak neliší od problému distribuce délky slov. Porovnejte své výsledky se závěry, ke kterým dospěl Best (2001). Porovnejte inventáře fonémů zkoumaných jazyků a pokuste se identifikovat možné Saportovy faktory.

Literatura

Best, K.-H. (2001). Zur Länge von Morphen in deutschen Texten. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt Verlag, 1–14.

Saporta, S. (1966). Phoneme distribution and language universals. In: Greenberg, J. H. (ed.), *Universals of language*. Cambridge: The MIT Press, 61–72.

6.3 DIVERZIFIKAČNÍ KONSTANTA

Hypotéza

Diverzifikaci daného jevu lze ve všech jazycích charakterizovat toutéž konstantou (Popescu, Altmann 2008).

Postup

Zaměřte se na diverzifikace řešené v části 7.5, „Distribuce diverzifikace". Podle dané hypotézy musejí všechny případy téhož jevu vykazovat velmi podobnou diverzifikaci, takže je možné určité vlastnosti rankové distribuce vystihující tuto diverzifikaci využít k její charakteristice. Popescu navrhl koeficient

$$c = \frac{R + f_{max} - f_{min} + 1 - L}{h} \ ,$$

kde R je počet diverzifikovaných tříd, tj. maximální rank, f_{max} a f_{min} jsou maximální a minimální frekvence distribuce, L je empirická délka křivky distribuce vypočítaná ze vzorce

$$L = \sum_{r=1}^{V-1} [(f_r - f_{r+1})^2 + 1]^{1/2} \ ,$$

představující součet euklidovských vzdáleností mezi sousedními frekvencemi a konečně h je h-bod, vypočítaný jako

$$h = \begin{cases} r, & \textit{pokud platí, že } r = f_r \\ \dfrac{f_1 r_2 - f_2 r_1}{r_2 - r_1 + f_1 - f_2} \ , & \textit{pokud neplatí, že } r = f_r \end{cases}$$

kde indexy 1 a 2 vyjadřují dvě libovolné sousedící třídy, např. $f_1 > r_1$ a $f_2 < r_1 + 1$. Pokud platí, že $f_{min} > R$, tj. nejnižší frekvence je větší než nejvyšší rank, od každé frekvence je možné odečítat $(f_{min} - 1)$. Zatímco délka křivky L se tím nezmění, bude možné snáze vypočítat h-bod.

Své výsledky porovnejte s níže uvedenou tabulkou (Popescu, Altmann 2008) a určete, zda leží v daném intervalu. Intervaly pro jednotlivá c jsou v tabulce uvedeny ve čtvrtém sloupci, intervaly pro průměrné

hodnoty *c* v pátém sloupci. Pokud vaše výsledky divergují, pokuste se najít faktor, který může být příčinou tohoto jevu.

TABULKA 6.3.1, Porovnání průměrných hodnot c
(Popescu, Altmann 2008)

Kategorie	***c***	s_c	***Int. c***	***Int. c***
Hlásky, fonémy, písmena	1,05	0,02	<1,00; 1,10>	<1,04; 1,06>
Slovní třídy (slovní druhy)	1,10	0,02	<1,06; 1,15>	<1,09; 1,12>
Rytmické vzorce	1,14	0,11	<0,92; 1,36>	<1,10; 1,18>
Paradigmatické třídy	1,15	0,05	<1,04; 1,26>	<1,09; 1,20>
Barevné třídy	1,18	0,07	<1,05; 1,32>	<1,15; 1,22>
Prepozice, postpozice, spojky	1,24	0,11	<1,03; 1,46>	<1,17; 1,32>
Diverzifikace pádů	1,33	–	–	–
Alomorfy plurálu	1,37	0,21	<0,97; 1,77>	<1,31; 1,43>
Afixy (významová diverzifikace)	1,39	0,16	<1,06; 1,71>	<1,32; 1,44>
Slova (významová diverzifikace)	1,47	0,21	<1,06; 1,88>	<1,44; 1,50>

Zaměřte se zejména na diverzifikaci funkcí a významů substantivních pádů, u nichž Popescu a Altmann uvedli pouze jednu položku, a určete možný interval.

Literatura

Best, K.-H. (2009). Diversifikation des Phonems /r/ im Deutschen. *Glottometrics 18,* 26–31.

Laufer, J., Nemcová, E. (2009). Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18,* 13–25.

Popescu, I.-I. Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18,* 32–39.

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17,* 97–111.

Rothe, U. (ed.) (1991). *Diversification process in language: grammar.* Hagen: Rottmann.

Sanada, H. (2009). *Diversification of postpositions in Japanese*.

6.4 SYNTETISMUS – ANALYTISMUS

Problém

Pouze na základě frekvencí slov z různých textů určete pozici vybraného jazyka na synteticko-analytické škále.

Postup

Syntetické jazyky mají mnoho slovních tvarů, které tak vyvářejí velmi dlouhou rankovou sekvenci. Analytické jazyky se oproti tomu vyznačují krátkou rankovou sekvencí slovních tvarů. Vypočtěte tedy rankovou distribuci slovních tvarů v textu (nebo několika textech v témže jazyce – nemíchejte mezi sebou různé texty) a následně vypočítejte hodnotu Popescova indikátoru

$$q = \frac{L_{max} - L}{N^{1/2}},$$

kde L je délka křivky mezi nejvyšší a nejnižší frekvencí vypočítaná jako součet euklidovských vzdáleností, L_{max} je maximální délka křivky a N je délka textu (počet slov – tokenů). Přesné definice uvádějí Popescu, Mačutek, Altmann (2009a, b).

Poté, co provedete analýzu několika textů v daném jazyce, vypočtěte průměrné q pro tyto texty a zjistěte, na které místo v níže uvedené tabulce váš jazyk patří. Pokud jste podrobili analýze nejméně 20 textů v některém jazyce uvedeném v tabulce a dospěli jste k hodnotě, která se výrazně liší od průměrného q, vypočtěte nevážený průměr své hodnoty a hodnoty v tabulce opravte. Pokud jste analyzovali nějaký nový jazyk, jednoduše tento jazyk a jeho hodnotu q doplňte do tabulky. Cílem je zpracovat tímto způsobem co nejvíce jazyků.

TABULKA 6.4.1, Průměrné hodnoty indikátoru *q*
(převzato z Popescu, Mačutek, Altmann 2009)

Jazyk	*q*	Jazyk	*q*
kannadština	0,273	ruština	0,382
latina	0,278	italština	0,412
maďarština	0,281	angličtina	0,435
indonéština	0,312	tagalština	0,446
maráthština	0,324	lakotština	0,449
němčina	0,334	maltština	0,479
čeština	0,336	markézština	0,504
rumunština	0,356	rarotongština	0,520

Jazyk	*q*	Jazyk	*q*
bulharština	0,370	havajština	0,542
slovinština	0,376	samojština	0,565

Literatura

Popescu, I.-I., Mačutek, J., Altmann, G. (2009a). A modified text indicator. In: Kelih, E., Levickij, V., Altmann, G., *Methods of text analysis*. Černivci: ČNU, 208–229.

Popescu, I.-I., Mačutek, J., Altmann, G. (2009b). *Aspects of word frequencies*. Lüdenscheid: RAM.

6.5 METODOLOGICKÉ PROBLÉMY

Problémy

(1) Je cílem typologie klasifikace? Pokud ano, je účelné považovat za základ takové klasifikace pouze gramatiku nebo pouze fonologii, případně kombinaci obojího?

(2) Je možné vyvodit nějaké závěry z typologické klasifikace jazyka, kterou znáte? Pokud ano, vezměte některé z těchto závěrů jako hypotézy a proveďte jejich empirické otestování. Pokud ne, jaký je vlastní cíl takové klasifikace?

(3) Je možné v typologii pracovat pouze s kategorickými (nominálními) pojmy nebo je účelnější (exaktnější, produktivnější) uplatňovat kvantitativní koncepty? Pokud dáváte přednost první variantě, jste si jisti, že byly vypořádány všechny možné nejasnosti? Jsou všechny jazyky jednoznačně přiřazeny k příslušným třídám? Pokud preferujete druhou variantu, pořiďte si přehled všech existujících typologických indikátorů, počínaje dílem Greenberga až po současnost.

(4) Všechny indikátory normalizujte, tj. transformujte je takovým způsobem, aby se pohybovaly v intervalu <0,1>. Vysvětlete, proč index, jehož horní interval odpovídá nekonečnu, neposkytuje žádný přijatelný popis jazykové reality. Existují v jazyce vlastnosti s nekonečnými hodnotami?

(5) Každý indikátor interpretujte. Je možné interpretovat index v intervalu $<0, \infty>$? Pokud ano, odpovídá hodnota 1 000 vysoké nebo nízké míře určité vlastnosti?

(6) Určete výběrové rozdělení indikátoru. Bude-li to nutné, poraďte se s odborníky na statistiku. Nebude-li se to dařit, vyřešte minimálně následující problém:

(7) Odvoďte teoretický odhad a rozptyl indikátorů a s jejich pomocí vytvořte asymptotický test za předpokladu normality. Stanovte intervaly spolehlivosti kolem průměrné hodnoty a na základě všech indikátorů proveďte předběžnou klasifikaci všech jazyků, které máte k dispozici. Dospějeme pomocí těchto indikátorů vždy k téže klasifikaci?

Literatura

Altmann, G., Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. München: Fink.

Anreiter, P. (1989). Transformierte sprachtypologische Profilvektoren. *Glottometrika 10*, 32–45.

Cysouw, M. (2005). Quantitative methods in typology. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative linguistics: an international handbook*. Berlin: de Gruyter, 554–557.

Fronzaroli, P. (1975). Problemi di classificazione delle lingue su base quantitative. In: *Colloquio sul tema: le tecniche di classificazione e loro applicazione linguistica*. Roma: Academia Nazionale dei Lincei, 123–141.

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics 26(3)*, 178–194.

Greenberg, J. H. (1974). *Language typology: a historical and analytical overview*. The Hague, Paris: Mouton.

Kasevič, V. S., Jachontov, S. (eds.) (1982). *Quantitative typology of Afro-Asiatic languages*. [v ruštině]. St. Petersburg: University Press.

Krámský, J. (1959). Quantitative typology of languages. *Language and Speech* 2, 72–85.

Kroeber, A. L. (1960). On typological indices 1. Ranking of languages. *International Journal of American Linguistics 26,* 171–177.

Kroeber, A. L. (1960). Three quantitative classifications of Romance. *Romance Philology 14,* 189–195.

Krupa, V. (1965). On quantification of typology. *Linguistics 12,* 31–36.

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

Lekomceva, M. I. (1963). K tipologii fonologičeskich struktur slova v slavjanskich jazykach. *Slavjanskoe jazykoznanie 1963*, 277–295.

Lekomceva, M. I. (1963). Tipologija fonologičeskich sistem. *Issledovanija po strukturnoj tipologii 1963,* 42–51.

Mejlach, M. (1973). Indeksy morfologičeskoj tipologii. In: *Problemy grammatičeskogo modelirovanija.* Moskva: Nauka, 155–170.

Sankaran, C. R., Taskar, A. D., Ganeshsundaruam, P. C. (1950). Quantitative classification of languages. *Bulletin of the Deccan College Institute 10,* 85–111.

Stepanov, A. V. (1995). Automatic typological analysis of Semitic morphology. *Journal of Quantitative Linguistics 2(2),* 141–150.

Winter, W. (1969–70). Some basic difficulties in the application of quantifying techniques to morphological typology. *Actes du X-e Congres International des Linguistes: 3.* Bucarest, 545–549.

6.6 SLOVOSLED (1)

Hypotéza

Ve studiích na téma jazykových univerzálií se můžeme setkat například s následující formulací: „Četnost, s jakou v jazycích s převažujícím slovosledem typu sloveso-podmět-předmět následuje adjektivum po substantivu, může být jen stěží dílem náhody". (Greenberg 1966: 85). Nahraďte tento vágní výrok výrokem exaktnějším.

Postup

Vyjděte z předpokladu, že v jazyce s převažujícím slovosledem typu sloveso-podmět-předmět existuje pravděpodobnost $p = 0{,}5$; že adjektivum bude stát za substantivem (a že pro $q = 1 - p = 0{,}5$ bude platit opačné uspořádání). V tomto případě bude distribuce počtu x jazyků v mezijazykovém výběrovém souboru o velikost n, u nichž bude slovosled predikován na základě náhody, odpovídat binomickému rozdělení s parametry $p = 0{,}5$ a n.

Navrhněte statistický test a stanovte metodu, na jejímž základě bude možné přesně určit hraniční frekvenci, po jejímž překonání by již tato hypotéza neměla platit.

Literatura

Greenberg, J. H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J. H. (ed.), *Universals of Language*. Cambridge, London: The MIT Press, 73–113.

6.7 SLOVOSLED (2)

Problém

Proveďte test Greenbergovy hypotézy z části 6.6, „Slovosled (1)".

Postup

Proveďte opětovnou analýzu mezijazykových (typologických) výběrových souborů. Stanovte pravděpodobnosti a proveďte testy významnosti. Podporují získaná data Greenbergovu hypotézu v její exaktní podobě?

Literatura

Greenberg, J. H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J. H. (ed.), *Universals of Language*. Cambridge, London: The MIT Press, 73–113.

6.8 SEKVENCE FONÉMŮ

Hypotéza

„V jazycích s výskytem odlučitelných i neodlučitelných vnitroslovních souhláskových shluků bude odlučitelná varianta významně četnější než neodlučitelná. (Odlučitelný shluk je definován jako sekvence, jejíž první část se nachází v koncové pozici a druhá část v počáteční pozici).“ (Saporta 1966: 67). Ověřte tuto hypotézu.

Postup

Potvrďte Saportovu hypotézu alespoň na jednom jazyce. Nejprve si pořiďte přehled všech vnitroslovních souhláskových shluků v daném jazyce. Rozlišujte dva případy: (a) slovníkový výskyt a (b) výskyt v textech, kde se může objevit více shluků v důsledku afixace nebo flexe. Následně proveďte test dané hypotézy s přihlédnutím ke dvěma různým alternativám: (1) Je odlučitelných shluků *více* než neodlučitelných a (2) mají vyšší *frekvenci*?

Existují zde ve skutečnosti čtyři hypotézy, přičemž každá z nich musí být řešena samostatně. U každé z nich proveďte statistický test, nerozhodujte se intuitivně.

Podle Saporty (1966) je příčinou tohoto jevu obecný princip jazykové ekonomie: „výskyt složité struktury implikuje výskyt jednodušší struktury". Rozveďte toto tvrzení a najděte další jevy, které odpovídají tomuto předpokladu.

Literatura

Saporta, S. (1966). Phoneme distribution and language universals. In: Greenberg, J. H. (ed.), *Universals of language* (2nd ed.). Cambridge: The MIT Press, 61–72.

6.9 SAPORTOVY SEKVENCE SOUHLÁSEK

Hypotéza

„Vyskytuje-li se C_1C_2-, pak je stejně pravděpodobný nebo ještě pravděpodobnější výskyt -C_2C_1" (Saporta 1966: 68), přičemž C = konsonant a shluky indikují pozici na začátku (C_1C_2-) nebo konci slova (-C_2C_1).

Postup

Nejprve najděte nějaká lingvistická východiska pro tuto hypotézu. Poté vyberte jazyk s velkým počtem souhláskových shluků, např. nějaký slovanský jazyk. Demonstrujte, že hypotéza nemusí platit. Upravte ji, definujte hraniční podmínky, další rysy daného jazyka apod. Jinými slovy, formulujte ji tak, aby byla přijatelná.

Literatura

Saporta, S. (1966). Phoneme distribution and language universals. In: Greenberg, J.H. (ed.), *Universals of language* (2nd ed.). Cambridge: MIT Press, 61–72.

6.10 FREKVENCE SLOV A ANALYTISMUS

Hypotéza

V textech ve výrazně analytických jazycích leží graf zipfovské mocninné funkce $f(r) = cr^{-a}$ *nad* hapax legomeny v rankové sekvenci slovních tvarů, zatímco ve výrazně syntetických jazycích leží *pod* nimi (Popescu, Mačutek, Altmann 2009: 104). Ověřte tuto hypotézu.

Postup

Určete rankovou distribuci slovních tvarů a textů z libovolného jazyka vyjma těch jazyků, jimiž se zabývají práce uvedené v seznamu referenční literatury, vypočtěte výše uvedenou zipfovskou funkci a aplikujte ji na absolutní frekvence. Tento efekt lze snadněji ukázat na datech z nějakého výrazně syntetického nebo výrazně analytického jazyka.

Míru analytismu/syntetismu lze vypočítat prostřednictvím následujícího indikátoru (Popescu, Mačutek, Altmann 2009: 106)

$$B = \frac{c}{(V - HL/2)^a} ,$$

kde a a c jsou parametry zipfovské funkce, které je nutné odhadnout z příslušných dat, V je velikost slovní zásoby textu (= počet různých slovních tvarů) a HL je počet hapax legomena. Na základě dat z různých jazyků odhadněte míru analytismu, vypočítejte jejich indikátory B a nakonec vypočítejte průměrnou hodnotu těchto B. Zařaďte zkoumaný jazyk na příslušné místo v tabulce 6.10.1.

TABULKA 6.10.1, Průměrné hodnoty indikátoru analytismu B u 20 jazyků (Popescu, Mačutek, Altmann 2009: 109)

	Jazyk	Průměrná hodnota *B*		Jazyk	Průměrná hodnota *B*
1	maďarština	0,2012	11	maráthština	12,302
2	čeština	0,7223	12	italština	12,787
3	latina	0,7982	13	lakotština	12,853
4	rumunština	0,8931	14	tagalština	13,913
5	němčina	0,9372	15	angličtina	14,514
6	slovinština	0,9418	16	markézština	18,108
7	kannadština	10,378	17	rarotongština	19,779
8	ruština	10,453	18	samojština	21,465
9	bulharština	10,495	19	maltština	21,861
10	indonéština	11,438	20	havajština	50,815

Literatura

Popescu, I.-I, Mačutek, J., Altmann, G. (2009). *Aspects of word frequencies*. Lüdenscheid: RAM.

7 Synergetika

7.1 FREKVENCE A POLYTEXTUALITA

Hypotéza

„… s rostoucí frekvencí slovesa klesá pravděpodobnost pevně stanoveného počtu jeho *argumentových struktur.*" (Thompson, Hopper 2001: 49).

Tato hypotéza je specifickou variantou obecnějšího tvrzení: s rostoucí frekvencí slova roste míra jeho výskytu v různých kontextech, tzn. kotextualita závisí na frekvenci (Köhler 1986).

Ověřte tuto hypotézu.

Postup

Existují tři možnosti, jak tuto hypotézu ověřit na datech z jakéhokoli jazyka:

(1) Pomocí frekvenčního slovníku vyberte 20 nejfrekventovanějších sloves (pro účely specifičtější hypotézy). Následně ve výkladovém slovníku spočítejte všechny fráze, idiomy apod., které jsou v něm u těchto sloves uvedeny.

(2) Najděte v nějakém textovém korpusu 20 nejfrekventovanějších sloves a u každého určete počet jeho různých argumentů.

(3) Z textů tvořících určitý korpus si zpracujte frekvenční seznam slov. Poté vyberte 100 nejfrekventovanějších slov a u každého určete počet jeho různých souvýskytů (typů kotextu). Zaměřte se na slovesa jako „typy", nikoli „tokeny", a zjistěte, které výrazy jim (a) předcházejí, (b) následují po nich, (c) oboje.

Ověřte na těchto datech hypotézu, že kotextualita (CT) je specifickou funkcí frekvence (F), tj.

$$CT = aF^b,$$

kde *a* a *b* jsou parametry. U Giesekinga (2002) a Köhlera (2002) byl tento vztah popsán opačně. Prokažte, že platí obousměrně. Zjistěte si parametry pro různé slovní třídy a pokuste se je příslušnými parametry charakterizovat.

Viz také Strauss et al. (2014, kap. „Délka slov a polytextualita" a „Kolokace").

Literatura

Gieseking, K. (2002). Untersuchungen zur Synergetik der englischen Lexik. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik*, 387–433.

Universitätsbibliothek Trier. [online]. Dostupné z: http://ubt.opus.hbz-nrw.de/volltexte/2004/279/ (cit. 24. března 2008).

Köhler, R. (1985). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch*. Berlin, New York: de Gruyter, 760–774.

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

Thompson, S. A., Hopper, P. J. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In: Bybee, J., Hopper, P., *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 27–60.

7.2 POLYSÉMIE A POLYTEXTUALITA

Hypotéza

Polytextualita roste s rostoucí polysémií podle mocninné funkce. Ověřte tuto hypotézu.

Postup

Polytextualitu je možné operacionalizovat různými způsoby v závislosti na zkoumaných jednotkách. Nejjednodušší operacionalizací z hlediska slov je počet různých textů v nějakém korpusu, v němž se dané slovo vyskytuje alespoň jednou. U morfémů či morfů je vhodným kvantifikačním nástrojem počet slovníkových výrazů obsahujících dané formy morfu/morfému. Obecně řečeno lze polytextualitu měřit počtem kotextů či kontextů určité jednotky.

Polysémii slov lze přibližně určit na základě počtu jejich jednotlivých významů uváděných výkladovým slovníkem, i když slovníkové rozlišování mezi homonymií a polysémií může být stejně problematické jako dílčí kategorizace významů. Použít lze rovněž elektronické verze slovníků. V případě morf(ém)ů lze polyfunkčnost určit buď na základě výhradně sémantických kritérií, nebo jejich kombinace s gramatickými prvky.

Jedním z výše uvedených způsobů si u zkoumaných jednotek opatřete hodnoty týkající se jejich polysémie (či polyfunkčnosti) a polytextuality. Roztřiďte tyto dvojice hodnot podle polysémie, tj. utvořte skupiny z dvojic hodnot s identickou polysémií. Vypočítejte průměrnou polytextualitu hodnot v jednotlivých skupinách. Výsledný počet průměrů bude odpovídat počtu různých hodnot polysémie. Dvojice < polysémie, průměrná polytextualita > představují data, na něž lze aplikovat odpovídající funkci.

Köhler (1985, 2005) odvozuje funkci

$$y = Ax^b$$

z diferenciální rovnice jako prvek synergetického řídicího cyklu. Aplikujte tuto funkci na svá data. Lepšího výsledku (určeného koeficientem determinace) lze případně dosáhnout použitím její rozšířené verze, pokud je řídící cyklus doplněn o další operátor (srov. Köhler 2006):

$$y = Ax^b e^{cx} \ .$$

Literatura

Köhler, R. (1985). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch*. Berlin, New York: de Gruyter, 760–774.

Köhler, R. (2006). Frequenz, Kontextualität und Länge von Wörtern – Eine Erweiterung des synergetisch-linguistischen Modells. In: Rapp, R., Sedlmeier, P., Zunker-Rapp, G. (eds.), *Perspectives on Cognition – A Festschrift for Manfred Wettler*. Lengerich: Pabst Science Publishers, 327–338.

7.3 DÉLKA MORFŮ A INVENTÁŘ FONÉMŮ

Hypotéza

„Průměrná délka morfů bude nepřímo úměrná počtu fonémů v inventáři." (Saporta 1966: 70). Ověřte tuto hypotézu.

Postup

Opatřete si data týkající se průměrné délky morfů a inventáře fonémů z několika jazyků. Nejsou-li k dispozici žádné frekvenční slovníky morfů, použijte alespoň jeden delší text v minimálně pěti jazycích. V ideálním případě by se mělo jednat o tentýž text, např. překlad nějakého krátkého textu. Nejprve najděte jednoduchou funkci, která by vyjadřovala tento vztah. Pokud aplikace funkce není uspokojivá (tj. koeficient determinace je nízký), přidávejte postupně další „kompenzační" faktory, jak to navrhují Jakobson a Gedney v poznámce k Saportovu článku, tj. pokuste se postihnout daný vztah tím, že přidáte další proměnné: počet kombinací fonémů uplatňujících se v daném jazyce, počet homonym a působení tónu anebo přízvuku. Sestavte řídící cyklus, v němž bude jako závislá proměnná figurovat průměrná délka morfů.

Literatura

Saporta, S. (1966). Phoneme distribution and language universals. In: Greenberg, J. H. (ed.), *Universals of language*. Cambridge: The MIT Press, 61–72.

7.4 FREKVENCE A POLYSÉMIE

Hypotéza

Existuje „… přímá souvislost mezi počtem různých významů slova a relativní frekvencí jeho výskytu." (Zipf 1945a: 144). Ověřte tuto hypotézu.

Postup

K ověření této hypotézy si pomocí frekvenčního slovníku nebo Wordnetu či jiných elektronických zdrojů sestavte výběrový soubor čítající minimálně 100 slov. V elektronických zdrojích a některých frekvenčních slovnících se uvádějí frekvence každého jednotlivého významů slova. Vy byste měli pracovat s celkovými frekvencemi slova (úhrnem frekvencí jednotlivých významů).

Pokud hypotéza platí, pak je závislost $S = f(F)$ vcelku jednoznačná. Někteří badatelé však varují před unáhleným zobecňováním a doporučují zohlednit různé hraniční podmínky (Ullmann 1966). V Köhlerově (2005) řídícím cyklu nenajdeme žádnou přímou spojitost mezi frekvencí a polysémií: jedná se o nepřímou relaci zprostředkovávanou délkou (srov. také Guiter 1974), kterou lze vyjádřit vzorcem

$$y = Ax^b \quad \text{nebo} \quad y = Ax^b e^{cx} \ .$$

Pomocí průměrů a vyhlazování nalezněte přímou závislost. Dosaďte do daného modelu Köhlerovy tzv. komunikační požadavky. Neomezujte se pouze na angličtinu. Zkoumejte tímto způsobem několik různých jazyků.

Tato souvislost může být eventuálně využita také pro účely stylistické analýzy. Zvolte si texty různého typu, spočítejte frekvence jednotlivých slov a pomocí výkladového slovníku určete u každého slova jeho polysémii. Vytvořte relaci $S = f(F)$, jež bude s největší pravděpodobností mocninnou funkcí, a zaměřte se podrobněji na parametry daných funkcí. Lze pozorovat rozdíly mezi texty různého typu, např. mezi vědeckými a poetickými texty, nebo lze pozorovat i stylistické rozdíly, např. mezi dvěma lyrickými básněmi?

Literatura

Carloni, F. (2000). Le relazioni statistiche tra frequenza e significato delle parole nella lingua italiana. *Italica 77(4)*, 523–534.

Guiter, H. (1974). Les relations fréquence-longueur-sens des mots (language romanes et anglais). *Atti del XII Congresso internazionale di linguistica e filologia romanza 14(4), 373–381. Napoli, 1970*, 15–20.

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

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch*. Berlin, New York: de Gruyter, 760–774.

Manin, D. Y. (2008). Zipf´s law and avoidance of excessive synonymy. *Cognitive Science 32(7)*, 1075–1098.

Ullmann, S. (1966). Semantic universals. In: Greenberg, J. H. (ed.), *Universals of language*. Cambridge: The MIT Press, 217–262.

Zipf, G. K. (1945a). The repetition of words, time-perspective and semantic balance. *The Journal of General Psychology 32*, 127–148.

Zipf, G. K. (1945b). The meaning-frequency relationship of words. *The Journal of General Psychology 33*, 251–256.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Reading, Mass: Addison-Wesley.

7.5 DISTRIBUCE DIVERZIFIKACE

Hypotéza

Dochází-li k diverzifikaci určité jednotky, frekvence jednotlivých prvků vykazuje pravidelnou distribuci.

Postup

Můžete si zvolit jakékoli jazykové jednotky, u nichž dochází k diverzifikaci, tj. prvky formální (fonetické, grafické), morfologické, sémantické, syntaktické, lexematické, dialektové, sociolektové apod. Ukázkovým příkladem je zkoumání významů spojky „a". Spočítejte frekvenci každého jejího jednotlivého významu zvlášť v daném textovém materiálu. Následně určete empirickou rankovou distribuci těchto jednotlivých významů. Identifikujte (a) funkci a (b) distribuci, pomocí níž ji můžete úspěšně modelovat.

V případě (a) aplikujte nejprve obvyklé zipfovské pojetí $f_r = c/r^a$ (r = rank, f_r = frekvence při ranku r) nebo Zipf-Mandelbrotovo pojetí $f_r = c/(r + b)^a$ a poté vyzkoušejte Popescovo pojetí $f_r = 1 + a \times exp(-r/b)$. Vypočítejte hodnotu koeficientu determinace určující shodu modelu s daty. V případě (b) nalezněte adekvátní distribuci. Začněte u Zipfa a Zipf-Mandelbrota (co se týká distribucí), poté pokračujte s geometrickým rozdělením podle Shentona a Skeese (srov. Shenton, Skees 1970; Wimmer, Altmann 1999; Mačutek 2008)

$$P_x = pq^{x-1}\left[1 + a\left(x - \frac{1}{p}\right)\right], \quad x = 1, 2, 3, \ldots,$$

kde a a p jsou parametry, $0 < p < 1$, $q = 1 - p$, $0 < a < 1/q - 1$. Proveďte test dobré shody pomocí chí-kvadrát testu. Najděte „nejlepší" model a analyzujte tímto způsobem všechny spojky.

Popište projevy tohoto druhu diverzifikace. U kterého modelu bylo dosaženo „nejlepší" shody? Existuje vzájemná souvislost mezi parametry jednotlivých modelů?

Věnujte pozornost také diverzifikaci jiných jevů (srov. Rothe 1991) a proveďte jejich porovnání s diverzifikací spojek.

Vysvětlete existenci pozorované pravidelnosti.

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.

Fan, F., Altmann, G. (2008). On meaning diversification in English. *Glottometrics 17,* 69–81.

Fan, F., Popescu, I.-I., Altmann, G. (2008). Arc length a meaning diversification in English. *Glottometrics 17,* 82–89.

Mačutek, J. (2008). On the distribution of graphemic representations. In: Altmann, G., Fan, F. (eds.), *Analyses of script. Properties and characters of writing systems.* Berlin, New York: de Gruyter, 75–78.

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17,* 97–111.

Popescu, I.-I., Altmann, G., Köhler, R. (2008). Zipf's law – another view. *Quality and Quantity.* [online].

Rothe, U. (ed.) (1991). *Diversification process in language: grammar.* Hagen: Rottmann.

Shenton, I. R., Skees, P. (1970). Some statistical aspects of amounts and duration of rainfall. In: Patil, G. P. (ed.), *Random counts in scientific work, Vol 3.* University Park: The Pennsylvania State University Press, 73–94.

Wimmer, G., Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

7.6 SYSTÉMOVÉ HRANICE A INTERAKCE

Problém

Žádnou jednotku (ať systémovou či nikoli) nelze konceptualizovat bez toho, aniž bychom uvažovali její hranice. V lingvistice jsou nám známy problémy související se segmentací mnoha jazykových jednotek (např. hlásek, slabik či nejrůznějších nespojitých jednotek), tj. s určením jejich hranic. O hranicích jazyků jako systémů toho zatím bylo řečeno velmi málo. Věnujte se podrobněji některým těmto hranicím. Uvažujte jazyk jako systém vyznačující se dalšími podsystémy, jednotkami, vlastnostmi.

Postup

Zamyslete se nad různými typy hranic pojícími se s konceptem jazyka. Existují hranice, které od sebe vzájemně oddělují různé jazyky (co například dialekty?), hranice uvnitř jazyka (mezi dílčími systémy, např. mezi syntaxí a morfologií, syntaxí a lexikem apod., mezi inventářem morf(ém)ů a lexémů nebo dalšími inventáři, mezi jednotlivými funkčními styly či sociolekty, mezi individuálními „jazyky"/idiolekty jednotlivce (s ohledem na kognitivní aspekt jazyka) atd. Uveďte další příklady částí jazyka a zkoumejte hranice mezi nimi.

Literatura

Než se tímto tématem začnete blíže zabývat, měli byste se obeznámit se základními pojmy a způsobem myšlení, které se pojí s moderní teorií systémů. Relevantní literaturu v tomto směru představují např.:

Altmann, G., Koch, W. (eds.) (1998). *Systems. New Paradigms for the Human Sciences*. Berlin: de Gruyter.

Bowler, T. D. (1981). *General Systems Thinking*. New York, Oxford: North Holland.

Bunge, M. (1979). *Treatise in Basic Philosophy, Vol. 4. Ontology II: A world of systems*. Dordrecht, Boston, London: Reidel.

7.7 JAZYK A TEXT

Problém

Co si myslíte o hranici a interakci mezi jazykem a textem?

Postup

Někteří lingvisté považují za systém nejen jazyk, ale také text. Určete nejprve rozdíl mezi těmito dvěma typy systémů (s ohledem na jejich dynamiku, funkci a strukturu). Poté sestrojte diagram znázorňující hranice, průnik a vzájemný vztah, resp. vztahy, mezi těmito dvěma systémy. Následně o svých zjištěních uvažujte na pozadí myšlenky, že (1) jazyk je sférou potenciality, zatímco text (parole) sférou reality, a že (2) jazyk je utvářen používáním při komunikaci a (3) jazyk je jen pouhý konstrukt, jenž slouží k popisu pravidelností jazykové komunikace.

Literatura

Bybee, J. (2006). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, J., Hopper, P. (2001). *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 1–26.

7.8 FREKVENCE A STÁŘÍ SLOV

Hypotéza

Čím je slovo starší, tím je frekventovanější. Ověřte tuto hypotézu.

Postup

S touto hypotézou se pojí dva problémy: (1) stáří slov nelze přesně určit. Pro naše účely jej lze odhadovat na základě roku/století, kdy se slovo poprvé objevilo

v písemných záznamech. (2) Slova také mohou zaniknout. Danou hypotézu je proto nutné specifikovat.

Stáří slova se dá nejsnadněji zjistit v historickém (etymologickém) slovníku. Pokud se uvádí pouze století, uvažujte vždy polovinu tohoto století. Z daného slovníku si pořiďte výběrový soubor nebo vyberte pouze slova jedné určité třídy. Na základě korpusu nebo frekvenčního slovníku následně určete frekvence daných slov. Prokažte, že mezi stářím a frekvencí existuje alespoň nějaká korelace. Proveďte tuto analýzu i na jiných jazycích než angličtině.

Pokud vaše data potvrzují výše uvedenou základní hypotézu, aplikujte na tuto hypotetickou závislost nějaký empirický vzorec. Vynechejte výrazy označující průmyslové produkty, neboť ty jsou mladé a frekventované. Pracujte vždy s průměry slov stejného stáří.

Literatura

žádná

7.9 DÉLKA A STÁŘÍ SLOV

Hypotéza

Čím je slovo starší, tím je kratší. Ověřte tuto hypotézu.

Postup

Pokud se slova v jazyce udrží po dlouhou dobu, pak musí být dostatečně frekventovaná. V takovém případě však dochází k jejich zkracování. Prostudujte jednotlivé soubory slov různých slovních tříd. Pokud to bude možné, zjistěte si jejich stáří stejným způsobem, jak je o tom pojednáno v části 7.8, „Frekvence a stáří slov". Pokuste se určit, u kterých tříd lze vysledovat platnost výše uvedené hypotézy, tj. specifikujte danou hypotézu. Zkoumejte tímto způsobem nějaký

jazyk s velkým množstvím nejednoslabičných slov. Vytvořte prototeorii souvislosti mezi zkracováním a stářím slov, tj. najděte odpovídající vzorce.

Literatura
žádná

7.10 VALENCE A POLYSÉMIE

Hypotéza
Čím větší je polysémie slova (slovesa, substantiva či adjektiva), tím větší je jeho valence.

Postup
Přestože valence může narůstat i bez nárůstu polysémie, lze předpokládat, že pokud dojde k nárůstu polysémie, vznikají nové případy valence. K ověření této hypotézy budete v první fázi potřebovat valenční slovník – který uvádí pouze malou podmnožinu lexika daného jazyka – a následně pak běžný výkladový slovník, v němž můžete dohledat významy slov. U každého vybraného slova určete obě hodnoty (valenci a polysémii) a najděte vzájemnou závislost (pro každý slovní druh zvlášť). Závislost nebude lineární. Poté odvoďte závislost z předpokladů nebo vyjděte ze synergetických úvah.

Literatura
žádná

7.11 DODATEK K SYNERGETICKÝM PROBLÉMŮM

Problém

Lingvistická literatura pojednává o množství diskurzivně-pragmatických funkcí slovosledu ve větách. Patří mezi ně například *určení východiska výpovědi*, *důraz*, *pojmová blízkost*, *aktualizace*, *jistota* nebo *naléhavost*. Sestavte synergetický model, který by znázornil slovosled jako multifunkční prostředek a zakomponujte do něj některé myšlenky týkající se funkčních ekvivalencí slovosledu ve vztahu k jednotlivým funkcím, o nichž byla zmínka výše.

Postup

Zpracujte přehled různých funkcí slovosledu popisovaných v jazykovědné literatuře. Najděte funkční ekvivalenty slovosledu, jež lze pozorovat v přirozených jazycích, a zamyslete se nad jejich specifickými výhodami a nevýhodami ve vztahu k jednotlivým funkcím. Spojte všechny tyto aspekty do synergetického modelu a odvoďte z něj alespoň jednu empiricky ověřitelnou hypotézu.

Literatura

Behaghel, O. (1932). *Deutsche Syntax. Eine geschichtliche Darstellung. Bd. IV.*, Heidelberg. [Germanische Bibliothek. I. Sammlung germanischer Elemenarund Handbücher. 1. Reihe: Grammatiken. 10. Bd].

Croft, W. (2003). *Typology and universals, 2nd edition*. Cambridge: Cambridge University Press.

Givón, T. (1984). *Syntax: A functional-typological introduction, Volume I.* Amsterdam: J. Benjamins.

Givón, T. (1990). *Syntax: A functional-typological introduction, Volume II.* Amsterdam: J. Benjamins.

Givón, T. (1988). The pragmatics of word order: predictability, importance and attention. Studies in syntactic typology. In: Hammond, M., Moravcsik, E., Wirth, J. (eds.), *Studies in syntactic typology,* Amsterdam: J. Benjamins, 243–284.

Greenberg, J. H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J. H. (ed.), *Universals of Grammar, 2nd edition*. Cambridge: The MIT Press, 73–113.

Haiman, J. (1983). Iconic and economic motivation. *Language 59,* 781–819.

Haiman, J. (1985). *Natural syntax*. Cambridge: Cambridge University Press.

Hawkins, J. A. (1983). *Word order universals*. New York: Academic Press.

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

Jespersen, O. (1942). *A Modern English Grammar on Historical Principles, IV*. Munksgaard: Copenhagen.

Jespersen, O. (1949). *A Modern English Grammar on Historical Principles, Part 2 (Syntax, Vol. 1)*. Copenhagen: Munksgaard, London: George Allen and Unwin.

Köhler, R. (1985). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch*. Berlin, New York: de Gruyter, 760–774.

Köhler, R. (2006). Frequenz, Kontextualität und Länge von Wörtern – Eine Erweiterung des synergetisch-linguistischen Modells. In: Rapp, R., Sedlmeier, P., Zunker-Rapp, G. (eds.), *Perspectives on Cognition – A Festschrift for Manfred Wettler*. Lengerich: Pabst Science Publishers, 327–338.

7.12 FONOTAKTIKA: OPTIMÁLNÍ VYUŽITÍ LINGVISTICKÉHO MATERIÁLU

Problém

Nedostatečné využívání materiálu (jinými slovy: potenciálních odlišností) je v jazyce známým faktem. Například k vytváření jednotek vyšší úrovně, např. morfů, využívají jazyky pouze malou část možných kombinací fonémů.

Totéž platí – ještě v markantnější podobě – pro využívání kombinací morfů při slovotvorbě.

Zdá se, že míra nedostatečného využívání potenciálu jazyka závisí na délce jednotek vyšší úrovně, tj. čím delší je řetězec sledovaných fonémů, tím menší počet permutací se při standardní (gramaticky správné) realizaci morfů vyskytuje.

Postup

Sestavte model, který bude obsahovat: (1) počet fonémů v inventáři daného jazyka, (2) délku morfů, (3) velikost inventáře morfů, (4) délku morfů (distribuci), (5) podobnost morfů, (6) redundanční komunikační požadavky a (7) komunikační požadavky jazykové ekonomie.

(1) Definujte odpovídajícím způsobem pojmy v bodech (1)–(7). Určitým způsobem zdůrazněte aspekt *podobnosti* (5) s ohledem na artikulační, sluchové a psycholingvistické faktory. Zaměřte se na vzájemné vztahy a vlivy mezi systémovými proměnnými a sestavte na jejich základě synergeticko-lingvistický model.

(2) Z daného modelu odvoďte jednotlivé ověřitelné hypotézy o distribuci vlastností systémových proměnných a o funkční závislosti mezi systémovými proměnnými. Na tomto základě vytipujte data, která by se dala využít pro účely empirických testů daných hypotéz, a následně tato data shromážděte. Proveďte příslušné statistické testy.

(3) Transponujte daný model z fonologické/morfologické roviny do roviny morfologické/lexikální a rozhodněte, jaké změny bude případně nutné provést.

(4) Po odpovídajících úpravách bude možné daný model aplikovat i na vyšší úrovně, např. syntaktickou. Jaké problémy bude třeba vyřešit?

Literatura

Kelih, E. (2009). Phonemverbindungen und Inventarumfang: Empirische Evidenz und Modellentwicklung. *Glottotheory 1(2),* 60–74.

Köhler, R. (1985). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch*. Berlin, New York: de Gruyter, 760–774.

Köhler, R. (2006). Frequenz, Kontextualität und Länge von Wörtern – Eine Erweiterung des synergetisch-linguistischen Modells. In: Rapp, R., Sedlmeier, P., Zunker-Rapp, G. (eds.), *Perspectives on Cognition – A Festschrift for Manfred Wettler.* Lengerich: Pabst Science Publishers, 327–338.

7.13 DÉLKA A POLYSÉMIE SLOV V ČÍNŠTINĚ

Problém

V synergetické lingvistice je vztah mezi délkou slova (WL) a polysémií (P) obvykle vyjádřen jako $WL = aP^{-b}$. Zjevně též platí i opačně, že $P = c(WL)^{-d}$. Je zde však drobný problém: polysémie nemůže být menší než 1 (pokud nebudeme vlastní jména považovat za slova s nulovým významem) a slovo obsahuje nejméně jednu slabiku. V některých slovanských jazycích sice existují neslabičné předložky, ty se však obvykle považují za proklitika. V obou relacích je tedy hodnota 1 asymptotou vztahů. Vyřešte alespoň jeden z následujících problémů: (1) opravte výše uvedené vzorce, (2) sestavte diferenciální rovnici a interpretujte ji, (3) aplikujte nový vzorec na níže uvedená data z čínštiny.

Postup

Problém (1) je jednoduchý. Doplnění hodnoty 1 do výše uvedených vzorců zajišťuje konvergenci k 1, když nezávislá proměnná konverguje k nekonečnu. Dostáváme tedy

(a) $WL = 1 + aP^{-b}$ $\qquad\qquad$ $P = 1 + c(WL)^{-d}$.

Nyní vytvořte diferenciální rovnici a interpretujte ji. Jedná se o nehomogenní diferenciální rovnici prvního řádu.

Breiterová (1994) uveřejnila následující data týkající se čínštiny (s použitím vážených průměrů polysémie):

TABULKA 7.13.1, Vztah délky slova a polysémie

Délka (*WL*)	Průměrná polysémie (*P*)
1	4,23
2	1,90
3	1,88
4	1,35

Ačkoli je tato řada krátká, aplikujte (b) $P = 1 + c\,(WL)^{-d}$ na tato data. Nemáte-li k dispozici dostatek tříd na test dobré shody, proveďte odhad c a d z prvních dvou tříd.

Literatura

Breiter, A. M. (1994). Length of Chinese words in relation to their other systemic features. *Journal of Quantitative Linguistics 1(3),* 224–231.

7.14 DÉLKA A FREKVENCE AFIXŮ

Hypotéza

V silně flektivních jazycích platí, že čím je afix delší, tím menší je jeho frekvence. Ověřte tuto hypotézu.

Postup

V tomto případě se zipfovský vztah ověřuje v obráceném pořadí.

K prvnímu testu můžete využít data ze španělštiny, která publikoval Urrea (2000: 111–112). Otestujte zvlášť prefixy (jež jsou většinou derivační) a sufixy (jež jsou většinou flektivní). Určete délku afixu (x) z hlediska počtu fonémů. Frekvenci lze určit jako průměrnou frekvenci (y) všech afixů téže délky. Navrhněte funkci jako model závislosti. Odůvodněte tento vztah.

Platí tento vztah také v silně aglutinačních jazycích? Pokud ne, proč?

Literatura

Urrea, A. M. (2000). Automatic discovery of affixes by means of a corpus: a catalog of Spanish affixes. *Journal of Quantitative Linguistics 7(2),* 97–114.

8 Filozofie vědy a obecné problémy

8.1 MÍRA KONSTITUENCE

Hypotézy

„… čím častěji se dva prvky vyskytují v řadě za sebou, tím pevnější bude jejich konstituentní struktura." (Bybee, Scheibman 1999; Bybee, Hopper 2001: 14). Ověřte tuto hypotézu.

Postup

První a nejdůležitější problém spočívá v koncipování postupu, jehož prostřednictvím by bylo možné zjistit míru pevnosti. Bez operacionalizace pojmu pevnosti a příslušného nástroje jejího měření nelze žádnou takovou hypotézu testovat. Je třeba rozlišovat mezi fonetickou a psanou formou textu, tj. je třeba vytvořit dvě různá kritéria pevnosti. Škálování by mělo být v maximální možné míře objektivní a foneticky uplatnitelné na všechny jazyky, pokud jde o psané formy, škálovaní by se mělo dát uplatnit na všechny jazyky, v nichž se užívá totéž písmo.

Na základě hodnot pevnosti a relativních frekvencí slov v korpusu lze provést analýzu závislosti. Nejslabší metodou je korelační analýza, neboť tou lze demonstrovat pouze lineární vztah. Vědecky nejkomplexnější způsob představuje teoretická derivace specifické hypotézy o příslušné formě závislosti a její matematické vyjádření, které lze následně ověřit na dostupných datech.

Jednou z možností je rozčlenit texty na morfy a vyčíslit jejich frekvence, pokuste se prokázat, že čím má morf vyšší frekvenci, tím větší je jeho odolnost vůči změnám (alomorfie), a čím častěji se dva morfy vyskytují spolu, tím vyšší je stupeň jejich vzájemné pevnosti.

Ohledně způsobu měření pevnosti se můžete inspirovat u Fana a Altmanna (2007), výpočty si ale adekvátně upravte.

Problém následujícím způsobem zobecněte: předložte argumenty pro tvrzení, že konstituence je spojitým jevem, tj. mezi frází, složeninou a odvozeninou neexistuje žádná jasná hranice, navzdory tomu, že školní gramatiky je takto jednoznačně vymezují, byť současně připouštějí existenci různých výjimek. Ve prospěch jednoho z argumentů může hovořit skutečnost, že ve výše formulované hypotéze se vyskytuje výraz „čím častěji", což nelze považovat za žádné jednoznačné vymezení.

Proveďte analýzu nejednoznačnosti hranic libovolných jazykových tříd a věnujte pozornost roli frekvence při tvoření těchto tříd.

Literatura

Boyland, J. T. (1996). *Morphosyntactic change in progress: a psycholinguistic approach*. Diss: Linguistic Department, University of California.

Bybee, J., Hopper, P. (2001). Introduction to frequency and the emergence of linguistic structure. In: Bybee, J., Hopper, P., *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: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative linguistics*. Černivci: RUTA, 177–189.

Krug, M. G. (2001). Frequency, iconicity, categorization: evidence from emerging modals. In: Bybee, J., Hopper, P., *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: J. Benjamins, 1–26.

8.2 CVIČENÍ Z FILOZOFIE VĚDY

„Na rozdíl od vrozených vzorců chování a na rozdíl od know-how je vědecké poznání veskrze konceptuální: tvoří jej systémy pojmů, mezi nimiž existují přesně vymezené vztahy."

(Bunge 2007: 51)

8.2.1 Pojem

(1) Zpracujte si seznam některých lingvistických pojmů a rozdělte je na (a) klasifikační (nominální škála), (b) ordinální (pořadová škála) a (c) metrické (intervalová nebo poměrová škála) pojmy. Zvolte si jeden nebo dva tyto pojmy a pokuste se je transformovat na pojmy (i) vyššího nebo (ii) nižšího řádu.

(2) Vychází klasifikace jazykových entit z předchozí znalosti základních vlastností zkoumaných objektů nebo předpokládaných tříd? Bude možné tyto základní vlastnosti pozorovat až na základě provedené klasifikace? Existují vůbec nějaké takové základní vlastnosti?

(3) Rozlišujte mezi pojmem *frekvence* a *obvyklost*, oba pojmy přesně definujte.

(4) Které z následujících lingvistických pojmů označují pozorovatelné jednotky nebo vlastnosti?

Slovo, slovní druh, morf(ém), frekvence, distribuce, typ fráze, slovesná valence, rod, délka, souvýskyt, závislost, důraz, ikonicita, parametr uspořádání, požadavek, inventář, produkční úsilí, příznakovost, přirozenost, projektivita, denotativní význam, konotace, text, jazyk.

(5) Které z těchto (nebo jiných) pojmů jsou intervenující pojmy (neobservační pojmy fungující jako prostředník mezi pojmy observačními) a které jsou hypotetickými konstrukty?

(6) Pokuste se zpřesnit definici (minimalizovat vágnost) nějakého zavedeného lingvistického pojmu (např. *lexikální kombinability*, srov. Levitskij [2005], *komplexity* nebo *ornamentality* písmových znaků nebo různých typů písma).

(7) Zamyslete se nad pojmem *fuzziness* v sémantice. Lze pojem *fuzzy význam* slova zpřesnit tím, že ho nahradíme pojmem pravděpodobnosti? Pokud dospějete k závěru, že tato možnost se nabízí jen v určitých případech, pokuste se takové případy charakterizovat. Zamyslete se nad důsledky takového nahrazení z hlediska intenze původního pojmu.

(8) Identifikujte observační pojem *fuzzy významu* (který by byl nezbytným předpokladem pro uplatnění postupu měření potřebného k empirickému určení hodnot funkce příslušnosti ve vztahu k *fuzzy významu* slova).

(9) Zamyslete se nad rozdíly mezi metaforami a importovanými pojmy. Jak budete v tomto ohledu posuzovat případy *mateřského uzlu*, *protojazyka*, *slovního pole*, *produkčního úsilí*, *jazykové ekonomie*, *entropie*?

(10) Zpracujte seznam základních pojmů vyjadřujících jazykové vlastnosti, tj. pojmů, které nejsou utvářeny pomocí jiných pojmů vyjadřujících jazykové vlastnosti. Totéž proveďte s jazykovými pojmy obecně.

Literatura

Bunge, M. (2007). *Philosophy of science. Vol. 1: From problem to theory*. New Brunswick, London: Transaction Publishers.

Levitskij, V. (2005). Lexikalische Kombinierbarkeit. In: Köhler, R. Altmann, G., Piotrovskij, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, 464–470.

8.2.2 Vědecké problémy

(11) Formulujte nějaký nový vědecký problém v oblasti nauky o jazyku a textu. Přesvědčte se, zda splňuje všechny formální a sémantické podmínky vědeckého problému (srov. Bunge 2005).

(12) Jak jste k novému problému dospěli? Kritikou známých řešení, zasazením známých řešení do nového kontextu, zobecněním vyřešeného problému nebo vztažením lingvistických idejí na pojmy z jiných oborů?

(13) Upřesněte logickou strukturu svého problému, zejména uveďte, která z veličin je neznámou proměnnou.

(14) Postupujte stejně jako v bodě 11 a odpovězte na následující otázky:
(a) Je váš mateřský jazyk typově syntetický?
(b) Jaké jsou vlastnosti frekvence slov?
(c) Existují nějaké jazyky, v nichž se nevyskytují žádná slovesa?
(d) Který z textů vašeho oblíbeného autora je nejkratší?
(e) Jak se tvoří kompozita ve francouzštině/ruštině?
(f) Proč aglutinační jazyky vykazují plošší TTR funkci než analytické?
(g) Zpracujte přehled některých kvantitativních vlastností vět.
(h) Jaká je průměrná délka maďarských slov?
(i) Jakým způsobem lze měřit syntaktickou nejednoznačnost?

(15) Jsou následující problémy dobře formulovány?
(a) Existuje nějaká vzájemná souvislost mezi délkou a valencí slov?
(b) Proč je jazyk systém?
(c) Jak lze vytvořit nástroj k měření estetické hodnoty?
(d) Který jazyk na světě má největší slovní zásobu?
(e) Je vždy možné určit, zda-li má symbol nějaký význam?
(f) Dochází v jazycích k neomezenému rozšiřování jejich lexika (případně jejich gramatiky)?

(16) Rozdělte problémy z části 14 a 15 (i další) na (a) empirické, (b) konceptuální, (c) metodologické a (d) hodnotící.

(17) Formulujte premisy související s vymezením problému v rámci bodu 10 nebo některého z bodů 13–15.

(18) Zamyslete se nad možností zobecnit vámi vymezený problém, přenést jej do jiného kontextu a uplatnit jej na jiný obor.

(19) Jak by mohlo řešení vašeho problému vypadat? Uveďte přibližnou charakteristiku typu odpovědi, která by daný problém řešila.

(20) Na základě nějakého učebního textu o dějinách lingvistiky pojmenujte některé historické problémy v tomto oboru. Uveďte stručně, jak byly tyto problémy případně vyřešeny.

(21) Jsou-li jazykové vlastnosti konceptuálními konstrukty, souhlasili byste s názorem, že jazyk má potenciálně nekonečné množství vlastností? Na čem bude jejich počet záležet?

(22) Je každou vlastnost možné měřit?

(23) Jsou-li vlastnosti konceptuálními konstrukty, jak je možné, že se mění? Co se mění?

(24) Existují v jazyce nějaké izolované vlastnosti?

(25) Odpovíme-li kladně na otázku v bodě 22, může nějaká jazyková vlastnost dosáhnout nekonečné hodnoty?

Literatura

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

Bunge, M. (2007). *Philosophy of science. Vol. 1: From problem to theory*. New Brunswick, London: Transaction Publishers.

Polya, G. (1957). *How to solve it*. New York: Doubleday Anchor Books.

8.3 RANKOVÁ FREKVENCE, OBECNÉ POJETÍ

Hypotéza

Pokud je určitá třída jazykových jednotek „správně" utvořena a jednotlivé prvky této třídy jsou seřazeny podle sestupné frekvence, pak tyto frekvence odpovídají funkci

$$f_r = 1 + a_1 \exp(-r/b_1) + a_2 \exp(-r/b_2) \text{ ,}$$

kde r = rank, f_r = frekvence při ranku r. Ověřte tuto hypotézu.

Postup

Tato hypotéza je zobecněním Popescu-Altmann-Köhlerova pojetí (2009), které se původně omezovalo jen na frekvence slov. Sestavte jakoukoli vhodnou skupinu jazykových jednotek, např. všech fonémů, slabik, názvů barev, zájmen, předložek, spojek, typů vedlejších vět (klauzí), typů kompozit, slovních tříd apod. Zjistěte frekvenci jejich výskytu v dlouhém textu, frekvence seřaďte podle ranku a aplikujte na získaná data výše uvedenou funkci. Při nízkém počtu ranků postačí první složka funkce. Při vysokých číslech ranků bude někdy nutné doplnit třetí exponenciální složku. Zaměřte se na chování funkce po přidání dalších složek.

TABULKA 8.3.1, Typy vedlejších vět u amerických spisovatelů
(Data převzata od Boyka 2005)

Typ vedlejší věty	Dreiser	Fitzgerald	Cronin	Steinbeck	Hemingway
podmětná	6	2	4	23	32
přísudková	5	5	2	13	4

Typ vedlejší věty	Dreiser	Fitzgerald	Cronin	Steinbeck	Hemingway
předmětná	647	306	246	173	208
přívlastková	488	235	194	165	121
časová	211	193	153	159	114
místní	37	15	21	26	26
příčinná	87	82	54	83	22
způsobová	141	87	50	63	33
důsledková	8	13	5	16	6
přípustková	41	12	46	16	9
účelová	12	3	2	10	3
podmínková	146	53	46	85	56

Jako příklad použijte Bojkova data (2005) týkající se typů vedlejších vět v dílech vybraných amerických spisovatelů (viz tabulka výše). Pro každého z autorů vytvořte rankovou sekvenci a aplikujte na ni výše uvedenou funkci. Použijte pouze jeden exponenciální komponent. Který z autorů se odlišuje od ostatních (porovnejte parametry b_1)?

Literatura

Boyko, J. (2005). Diferencijni parametri rečenija jak determinanta avtors'kogo stilju. In: Altmann, G., Levickij, V., Perebyjnis, V. (eds.), *Problems of Quantitative Linguistics*. Černivci: RUTA, 292–305.

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1.* Olomouc: Univerzita Palackého v Olomouci.

Popescu, I.-I., Altmann, G., Köhler, R. (2009). Zipf's law – another view. *Quality and Quantity.* [online], cit. 9. května 2009.

8.4 UNIVERZÁLIE, ZÁKONY A TEORIE

Problém

(a) Jsou „jazykové univerzálie" zákony? (b) Jsou „gramatické teorie" teoriemi? Pokuste se odpovědět na tyto otázky.

Postup

(1) Zaměřte svou pozornost na pojem jazykových univerzálií, jak o nich pojednává Greenberg (1978). Naplňují tvrzení tohoto druhu podmínky zákonů (jak je uvádí Bunge [1967])? Pokud dospějete k závěru, že se nejedná o zákony, co jiného to je?

(2) Jsou teorie X-bar, HPSG a další „gramatické teorie" systémy univerzálních zákonů, které vysvětlují pozorované skutečnosti? Pokud dospějete ke kladné odpovědi, explicitně některé z těchto zákonitostí formulujte a ověřte, zda skutečně platí jako zákony.

(3) Problém můžeme postavit také opačně: jsou lingvistické zákony jazykovými univerzáliemi?

Literatura

Bunge, M. (1967). *Scientific research I, II.* Berlin: Springer.

Cysouw, M. (2005). *Quantitative methods in typology.* In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistics. An International Handbook.* Berlin: de Gruyter, 554–578.

Greenberg, J. H. (ed.) (1975). *Universals of Human Language, Vol 1: Method and Theory*. Stanford: Stanford University Press.

8.5 POZOROVATELNOST

Problém

Určete pozorovatelnost jazykových jednotek.

Postup

Vytvořte si seznam jazykových entit, které vás zajímají (jednotky, vlastnosti, systémy apod.), a vyhodnoťte jejich pozorovatelnost. Je možné rozlišovat mezi přímým a nepřímým pozorováním. Někteří filozofové však definují pozorovatelnost výhradně s ohledem na lidské smysly, tj. za pozorování jako takové nepovažují pozorování prostřednictvím nástrojů (srov. van Fraassen 1980).

(1) Které jazykové jednotky lze pozorovat přímo, které nikoli?

(2) Které nástroje používají lingvisté k nepřímému pozorování?

(3) Je počítání metodou přímého pozorování? Pokud ne, jaké nástroje se používají k zjišťování počtu? Souhlasí vaše závěry s názorem, že smyslové vnímání je přímým pozorováním (používáme svůj kognitivní aparát také k tomu, abychom v tom, co vidíme nebo slyšíme, rozpoznali určitou charakteristiku známého prvku)? Je nějaký rozdíl v tom, když k počítání použijeme jen papír a tužku?

(4) Objasněte úlohu dichotomie *langue–parole* ve vztahu k otázce pozorovatelnosti.

(5) Zamyslete se nad stanoviskem, které by ve vztahu k dichotomii *langue–parole* museli zaujmout realisté a antirealisté.

Literatura

Van Fraassen, B. C. (1980). *The scientific image*. Oxford, New York: Oxford University Press.

9 Různé problémy

9.1 DÉLKA KŘIVKY A EVOLUCE JAZYKA

Problém

Typologickou evoluci jazyka lze zpětně vysledovat pomocí délky křivky rankové distribuce slovních tvarů.

Postup

Postupujte stejně jako v části 6.1, „Délka křivky a typologie", ale tentokrát analyzujte texty ve zvoleném jazyce z různých století. Sledujte změnu parametru *b* a určete, zda se daný jazyk vyvíjí spíše k syntetičtějšímu nebo analytičtějšímu typu.

Předveďte, které románské jazyky vykazují největší tendenci k analytismu. V poslední době se vedou spory o možný vývoj němčiny směrem k analytismu. Prověřte tuto otázku.

Zaměřte se na některé indonéské, melanéské a polynéské jazyky a demonstrujte geografické rozdělení analytismu v rodině austronéských jazyků. Tento problém lze vyřešit i bez znalosti těchto jazyků, neboť stačí spočítat slovní tvary, texty v těchto jazycích lze jednoduše nalézt na internetu.

Demonstrujte vývoj slovanských jazyků a geografické rozdělení syntetismu v Evropě. Lze dospět k závěru o existenci oblastních vlivů? Doporučili byste tuto metodu využívat v rámci dialektologie?

Literatura

Popescu, I.-I., Mačutek, J., Altmann, G. (2008). Word frequency and arc length. *Glottometrics 17,* 18–44.

9.2 ZDVOŘILOST

Problém

Zaměřte se na některé vlastnosti „zdvořilých" slov a výrazů a porovnejte je s „neutrálními" slovy.

Postup

„Zdvořilá" slova nebo výrazy se vyznačují určitými charakteristikami, které je odlišují od každodenních (neutrálních) výrazů. Vytvořte si soubor zdvořilých slov nebo výrazů a porovnejte je z různého hlediska (prozodického, fonologického, morfologického, lexikálního, sémantického apod.) s jejich „normálními" protějšky. Proveďte toto porovnání kvantitativně, tj. kvantifikujte dané vlastnosti a rozdíly mezi nimi. Demonstrujte existenci různých stupňů zdvořilosti a vyjádřete je na kvantitativní škále. Věnujte pozornost rovněž nezdvořilosti a jejím vlastnostem.

Zaměřte se z tohoto pohledu na jazyky jihovýchodní Asie. V případě japonštiny je možné najít již zpracované seznamy výrazů s odstupňovanou mírou zdvořilosti. Postačí však, pokud své probandy vyzvete, aby formulovali určitou otázku s různě odstupňovanou mírou zdvořilosti, a po jiných osobách budete chtít, aby následně provedly intuitivní škálování rozdílů v takto položených otázkách.

Literatura

Altmann, G., Riška, A. (1966). Towards a typology of courtesy in language. *Anthropological Linguistics 8,* 1–10.

Beeching, K. (2002). *Gender, politeness and pragmatic particles in French*. Amsterdam: J. Benjamins.

Brown, P., Levinson, S. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Ide, S. (1989). Formal forms and discernment: two neglected aspects of universals of linguistic politeness. *Multilingua 8(2/3)*, 223–248.

Jemmy, H. (2007). *What is politeness? I've never heard of it before, can I put it in my mouth?* Wigan: Pieperback Books.

Journal of Politeness Research. Language, Behaviour, Culture. Berlin, New York: de Gruyter.

Lakoff, R. (1975). *Language and woman's place*. New York: Harper & Row.

Matsumoto, Y. (1988). Reexamination of the universality of face: politeness phenomena in Japanese. *Journal of Pragmatics 12,* 403–426.

Mills, S. (2003). *Gender and politeness*. Cambridge: Cambridge University Press.

Stadler, S. A. (2007). *Multimodal (im)politeness. The verbal, prosodic and nonverbal realization of disagreement in German and New Zealand English*. Hamburg: Kovac.

Watts, R. J. (2003). *Politeness*. Cambridge: Cambridge University Press.

Watts, R. J., Ide, S., Ehlich, K. (eds.) (2006). *Politeness in language. Studies in its history, theory and practice*. Berlin, New York: de Gruyter.

9.3 DISTRIBUCE SLOVNÍCH TŘÍD V PŘÍSLOVÍCH

Problém

Definice pojmu přísloví (např. „Není všechno zlato, co se třpytí.“) obsahuje minimálně formální a pragmatickou část, tj. (1) přísloví vždy sestává z úplné věty (představuje propozici) a (2) je všeobecně známé a užívané.

Zkoumejte frekvenční distribuci slovních druhů vyskytujících se v příslovích. Výsledky porovnejte s jinými druhy jazykového materiálu.

Postup

Každému slovu v souboru přísloví přiřaďte tag slovního druhu, tyto tagy spočítejte a získaná data následně uspořádejte ve formě rankové distribuce. Určete, která teoretická rozdělení pravděpodobnosti odpovídají těmto datům (lze očekávat

jednu z „distribucí diverzifikace“ nebo např. Zipfovo useknuté rozdělení zeta $P_x = C/x^a, x = 1, 2, 3, \ldots, x_{max}$). Své zjištění zdůvodněte.

Literatura

Grzybek, P. (2004). A quantitative approach to lexical structure of proverbs. *Journal of Quantitative Linguistics 11/1–2,* 79–92.

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18,* 32–39.

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17,* 97–111.

Rothe, U. (ed.) (1991). *Diversification process in language: grammar.* Hagen: Rottmann.

9.4 KÖHLEROVY MOTIVY V PŘÍSLOVÍCH

Problém

Zkoumejte sekvence (neboli „motivy“) délky, frekvence, polysémie, polytextuality ve vztahu k jejich rankové, délkové a jiné distribuci. Dá se říci, že přísloví vykazují v tomto smyslu pravidelnější struktury než jiný jazykový materiál?

Postup

Za slova, morfy, slabiky atd. v příslovích dosaďte hodnoty výše zmíněných proměnných. Podle metody, kterou uvádějí Köhler (2006) a Köhler, Naumann (2008), vytvořte motivy/sekvence a určete, zda jednotlivé typy rozdělení potvrzují zjištění popsaná v literatuře. Lze mezi parametry rozdělení a parametry jiného materiálu pozorovat významné rozdíly?

Literatura

Grzybek, P. (2004). A quantitative approach to lexical structure of proverbs. *Journal of Quantitative Linguistics 11/1–2,* 79–92.

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: Preisach, B., Schmidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications. Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg.* Berlin, Heidelberg: Springer, 637–646.

9.5 SÉMANTICKÉ ROLE V PŘÍSLOVÍCH

Hypotéza

Sémantické role vykazují v příslovích pevné distribuční vzorce.

Postup

(1) Přiřaďte syntaktickým konstituentům v souboru přísloví nějaké sémantické role (např. AGENS, OBJEKT, INSTRUMENT) a stanovte jejich frekvenční rozdělení.

(2) Každé přísloví lze popsat prostřednictvím určitého schématu rolí (např. AGENS-OBJEKT-INSTRUMENT). Spočítejte frekvence těchto schémat ve vašem souboru a určete jejich rozdělení.

(3) Svá zjištění se pokuste interpretovat na pozadí synergeticko-lingvistického pojetí funkce přísloví. (Nápověda: uvažujte přísloví jako vysvětlení každodenních situací zprostředkovaná jazykovým kódem.)

Literatura

žádná

9.6 POČET A DÉLKA PŘÍSLOVÍ

Problém

Lze očekávat nějakou souvislost mezi počtem přísloví v určitém jazyce (souboru) a jejich průměrnou délkou? Pokud ano, opodstatněte takový názor.

Postup

Vytvořte distribuci délek přísloví v souboru a formulujte nějaký závěr. Délku určete dvojím způsobem: podle počtu slov a podle počtu klauzí.

Literatura

žádná

9.7 VĚTNÉ STRUKTURY V PŘÍSLOVÍCH

Hypotéza

U přísloví existuje zvláštní, silně asymetrická ranková distribuce typů větné struktury. Ověřte tuto hypotézu.

Postup

Analyzujte věty v souboru přísloví, co se týká jejich struktury z hlediska hlavních a vedlejších vět (klauzí), viz Tab. 9.7.1.

TABULKA 9.7.1, Syntaktická struktura přísloví z hlediska hlavních a vedlejších vět

Hlad je nejlepší kuchař.	H	hlavní věta
Jak si kdo ustele, tak si lehne.	V_{PuZ}+H	vedlejší věta příslovečná způsobová + hlavní věta

Neříkej hop, dokud jsi nepřeskočil.	$H+V_{PuČ}$	hlavní věta + vedlejší věta příslovečná časová

Stanovte rankovou distribuci větných vzorců a určete odpovídající teoretické rozdělení frekvence. Výsledek interpretujte na pozadí funkce přísloví v komunikaci.

Literatura
žádná

9.8 IDENTIFIKACE VARIANT FRAZEOLOGICKÝCH PRVKŮ

Problém
Frazeologické prvky (např. přísloví, fráze, rčení, „citace“) se často užívají (ať už vědomě či nevědomě) nesprávným nebo pozměněným způsobem. Jaké podobnosti/rozdílnosti mezi reálně užívanou variantou a její původní formou lze postulovat? Jak lze tyto podobnosti měřit? Jak velká musí být podobnost takové varianty, aby ji bylo možné rozpoznat?

Postup
Z novin a dalších zdrojů (variantami frází, názvů knih nebo filmů jsou často novinové titulky, varianty přísloví někdy obsahují také různé sbírky přísloví a také mezilidská komunikace může být v tomto směru bohatým zdrojem) si shromážděte potřebný materiál. Určete druhy (fonologické, morfematické, lexikální, syntaktické atd.) rozdílností v daném materiálu. Definujte způsoby určování podobnosti/odlišnosti, např. na základě Levenshteinovy editační vzdálenosti. Zamyslete se nad možností různých způsobů kombinovaní jednotlivých druhů podobnosti (např. pomocí vícerozměrného vektoru podobnosti).

Literatura

Gonzalo Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys 33(1),* 31–88.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. In: *Doklady Akademii Nauk SSSR 163(4),* 845–848. [v ruštině]. [Vyšlo v angličtině jako: *Soviet Physics Doklady 10(8) (1966),* 707–710.]

9.9 SYNONYMIE A (NE)ZDVOŘILOST

Problém

Čím je slovo méně zdvořilé, tím více má synonym. Ověřte tuto hypotézu.

Postup

Zvolte si buď nějaký slovník slangu nebo jazyk, v němž se míra zdvořilosti rozlišuje lexikálními nebo gramatickými prostředky, např. javánštinu nebo japonštinu. Pořiďte si seznam nezdvořilých (tvarů) slov a jejich neutrálních nebo zdvořilých ekvivalentů. Odstupňujte jednotlivá slova podle míry nezdvořilosti, dohledejte příslušná synonyma a po provedeném sčítání vyjádřete danou hypotézu formálním způsobem.

Literatura

žádná

9.10 PROCES ZÁNIKU V DIALEKTOLOGII

Hypotéza

Čím jsou si dvě lokality (místa) v rámci oblasti téhož jazyka vzdálenější, tím větší jsou dialektové rozdíly mezi nimi.

Postup

Navrhněte jednoduchý stochastický proces zániku znázorňující úbytek podobnosti (fonetické, lexikální atd.) v závislosti na rostoucí geografické vzdálenosti mezi dialekty. Použijte Poissonův proces, v němž čas nahradíte vzdáleností. Do jednotlivých vzorců zapracujte parametr představující určité hraniční podmínky (např. přirozené hranice, špatnou komunikaci apod.). Danou hypotézu ověřte pomocí dostatečného množství dat z dialektometrie.

Alternativní postup: použijte teorii difuze z biologie, sociologie atd.

K měření podobnosti uplatněte různé metody. Levenshteinova vzdálenost je spíše orientačním měřítkem, jehož prostřednictvím nelze vyjádřit jemnější fonetické rozdíly. Prostudujte si rovněž starší literaturu.

Literatura

Goebl, H. (1982). *Dialektometrie. Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. [Denkschriften der Österreichischen Akademie der Wissenschaften, phil.-hist. Klasse, 157], Wien.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Thesis, Rijksuniversiteit Groningen.

Nerbonne, J., Heeringa, W., Kleiweg, P. (1999). Edit distance and dialect proximity. In: Sankoff, D., Kruskal, J. (eds), *Time warps, string edits and macromolecules: The theory and practice of sequence comparison*: v–xv. Stanford: CSLI Press.

9.11 MOTIVY DÉLKY

Problém

Určete předpokládanou délku motivů délky a porovnejte ji s odpovídající empirickou průměrnou délkou. Následně proveďte totéž u délky motivů frekvence a polysémie.

Postup

Všechny druhy motivů, jak je definuje Köhler (2006), jakož i F-motivy, které již dříve definoval Boroda (1982) pro hudbu, mají v praxi omezenou délku. Jelikož motiv je podle definice sekvencí monotónně klesajících (nebo rostoucích) hodnot, můžeme danou situaci považovat za sekvenci binárních jevů: na každé pozici v rámci sekvence sledovaných hodnot existuje pravděpodobnost p, že daná hodnota je menší nebo rovna předcházející hodnotě (což by vedlo k nárůstu délky aktuálního motivu o jednotku), a pravděpodobnost $q = 1 - p$, že hodnota na dané pozici je větší než předcházející hodnota (což by vedlo k ukončení aktuálního motivu a přivodilo zahájení motivu nového). Předpokládanou délku motivu je tudíž možné určit prostřednictvím geometrického rozdělení. Pravděpodobnost pro motiv délky x je proto

$$P(L = x) = qp^{x-1}, \; x = 1, 2, 3, \ldots$$

Ze získaných dat vypočtěte p a q na základě zjištění relativního počtu změn vašich hodnot na všech pozicích v sekvenci na menší nebo stejné hodnoty ($= \hat{p}$). Toto číslo současně použijte k odhadu pravděpodobnosti p. Ověřte, zda se geometrické rozdělení hodí pro modelování vašich dat.

Věnujte rovněž pozornost některým problémům týkajícím se motivů, o nichž je pojednáno u Strausse et al. (2014) a v této kapitole.

Literatura

Boroda, M. G. (1982). Häufigkeitsstrukturen musikalischer Texte. In: Orlov, J. K., Boroda, M. G., Nadarejšvili, I. Š. (eds.), *Sprache, Text, Kunst. Quantitative Analysen*. Bochum: Brockmeyer, 231–262.

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: Preisach, B., Schmidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications. Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg.* Berlin, Heidelberg: Springer, 637–646.

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

9.12 FREKVENCE A PRODUKČNÍ ÚSILÍ (POKRAČOVÁNÍ)

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). Ověřte tuto hypotézu na různých případech.

Postup

Strauss et al. (2014, kap. „Frekvence a produkční úsilí") hovoří o významu, záběru, formulaci a dalších obecných aspektech této hypotézy. Pokud jste již vyřešili tyto počáteční problémy, zvolte si nějaký různě realizovaný jazykový jev a prověřte platnost dané hypotézy. Laufer (2009) například zkoumal způsoby vyjádření slovesného vidu v němčině. Najděte další jevy a na základě autentického jazykového úzu tuto hypotézu konkretizujte a zpřesněte.

Literatura

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

Laufer, J. (2009). [osobní sdělení].

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

9.13 FOURIEROVA ANALÝZA

Problém

Najděte v textech 10 různých sekvencí, které mohou vykazovat určité projevy cyklického opakování, „pravidelného kmitání", (princip vlnového pohybu). Pomocí Fourierovy analýzy je formálně popište.

Postup

Proveďte analýzu jednoho textu a zaměřte se na všechny dobře definovatelné sekvence, nebo si zvolte jen jeden typ sekvence a zkoumejte jej v rámci většího počtu textů. Pak svůj výsledek buď zobecněte na popis sekvencí na různých rovinách jazyka, nebo na popis chování jednoho typu sekvence.

Základní praktické informace o Fourierově analýze lze najít v učebních textech pojednávajících o problematice časových řad, jednoduchý návod poskytuje také Altmann (1988: 197). Komplexní statistické softwarové balíky nabízejí funkce, které Fourierovu analýzu provedou automaticky.

Mezi příklady cyklického opakování patří: (a) počet daktylů v sekvenci veršů, (b) sekvence slovních či větných délek v textu, (c) pozice důrazů na slovech tvořící binární sekvenci 10010110…, (d) sekvence vzdáleností mezi stejnými prvky atd.

Omezte počet koeficientů na maximální možnou míru. Svůj postup a výsledky odůvodněte. Oscilaci můžete také zachytit pomocí diferenčních rovnic, jejichž prostřednictvím lze zároveň prokázat, zda sekvence téhož druhu vykazují stejný řád ve všech textech. Ukažte, které (sekvenčně prezentované) vlastnosti jsou nižšího řádu a které vyššího řádu. Výsledek interpretujte a zasaďte do lingvistického kontextu.

Literatura

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

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

Hřebíček, L., Altmann, G. (1996). The levels of order in language. *Glottometrika 15,* 38–61.

10 Pragmatika

10.1 FREKVENČNÍ DISTRIBUCE MLUVNÍCH AKTŮ

Problém

Mluvní akty lze považovat za jazykové jednotky podobně jako slova, slabiky, fráze, věty, hreby atd. Musejí tudíž podléhat určitým pravidelnostem. Identifikujte některé z těchto pravidelností.

Postup

Vyjděte z některé existující klasifikace mluvních aktů. Klasifikace ani jiné deskriptivní prostředky nejsou nositeli pravdivostních hodnot, neexistuje tudíž žádná „správná" klasifikace, ale pouze více či méně vhodná (vzhledem k danému účelu). Zkuste různé klasifikace a vyhodnoťte, která z nich nejvíce odpovídá kvantitativně vyjádřeným pravidelnostem. Podívejte se na tuto klasifikaci: Bach, K. *Routledge Encyclopedia of Philosophy.* [online]. Dostupné z: http://online.sfsu.edu/kbach/spchacts.html (cit. 15. května 2009).

Konstativy: *ujištění, tvrzení, oznámení, odpovědi, označování, prohlášení, klasifikování, shoda, potvrzení, dohady, popření, nesouhlas, prozrazení, zpochybňování, identifikování, informování, trvání na něčem, předpovídání, hierarchizování, referování, konstatování, ujednání.*

Direktivy: *rady, upozornění, dotazy, prosby, zamítnutí, výmluvy, zákazy, pokyny, příkazy, svolení, žádosti, požadavky, návrhy, pobídky, varování.*

Komisivy: *souhlasy, záruky, pozvání, nabídky, sliby, přísahy, nabízení pomoci.*

Akceptivy: *omluvy, kondolence, gratulace, pozdravy, děkování, akceptování (braní na vědomí a uznávání).*

Proveďte transkripci nebo anotaci nějakého dramatu a transkribujte alespoň jedno jednání nějakého dramatu s ohledem na mluvní akty. Připravte dvě verze: (1) s rozlišením jednotlivých postav dramatu, (2) s textem jako celkem.

Vypočtěte frekvence jednotlivých mluvních aktů (připravíte-li si soubor obsahující jen sekvence tagů mluvních aktů, můžete použít nějaký vhodný software). Verzi (2) lze získat z verze (1) prostým součtem. Zpracujte rankové pořadí frekvencí pro každou osobu zvlášť i pro text jako celek (bez rozlišování jednotlivých osob).

(1) Demonstrujte, že ranková distribuce mluvních aktů sleduje jednu z těchto funkcí:

(a) $f(r) = ar^{-b}$

(b) $f(r) = 1 + ae^{-br}$

(2) Demonstrujte rozdílnost parametrů *a* a *b* u jednotlivých postav. Existuje nějaká korelace mezi těmito parametry a určitou vlastností dané postavy (např. dominancí, servilitou, nervozitou apod.)? Také vlastnosti postav by měly být samozřejmě kvantifikovány, třebaže pouze na ordinální škále, např. „dominance" v rozmezí od 0 – nulová dominance – po 10 – silná dominance.

(3) Vypočtěte průměr, rozptyl a třetí centrální moment rankové distribuce každé postavy a zaneste je do Ordova schématu (srov. Strauss et al. 2014, kap. „Ordovo kritérium"). Jak byste charakterizovali jednotlivé postavy ve vztahu k jejich pozici v Ordově schématu?

(4) Pokud zpracujete nějaké drama jako celek, výstupy rozdělte na postavy a jednání. Sledujte pohyb každé postavy během jednotlivých jednání v rámci Ordova schématu.

(5) Vypočtěte rankovou sekvenci a zaneste Ordovy funkce do Ordova schématu pro každé jednání zvlášť (u postav toto dělení neprovádějte). Pozorujete nějaký posun mezi začátkem a koncem dramatu?

Literatura

Alston, W. P. (2000). *Illocutionary acts and sentence meaning*. Ithaca: Cornell University Press.

Austin, J. L. (1962). *How to do things with words*. Cambridge: Harvard University Press.

Bach, K., Harnish, R. M. (1979). *Linguistic communication and speech acts*. Cambridge: The MIT Press.

Cohen, A. D. (1996). Speech acts. In: McKay, S. L., Hornberger, N. H. (eds.), *Sociolinguistics and language teaching.* Cambridge: Cambridge University Press, 383–420.

Doerge, F. C. (2006). *Illocutionary acts – Austin´s account and what Searle made out of it.* [online]. Dostupné z: http://tobias-lib.ub.uni-tuebingen.de//volltexte/2006/2273/

Grice, H. P. (1989). *Studies in the way of words*. Cambridge: Harvard University Press.

Olshtain, E., Cohen, A. D. (1989). Speech act behavior across languages. In: Dechert, H. W. et al. (eds.), *Transfer in production.* Norwood, NJ: Ablex, 53–67.

Sander, Th. (2002). *Redesequenzen. Untersuchungen zur Grammatik von Diskursen und Texten*. Paderborn: mentis.

Searle, J. (1969). *Speech acts: an essay in the philosophy of language*. Cambridge: Cambridge University Press.

Searle, J. (1975). Speech acts. In: Cole, P., Morgan, J. L. (eds.), *Syntax and semantics, 3: Speech acts: 59–82*. New York: Academic Press. [Přetištěno v: Davis, S. (ed.) (1991), *Pragmatics: A reader.* Oxford: Oxford University Press, 265–277.]

Staffeldt, S. (2008). *Einführung in die Sprechakttheorie. Ein Leitfaden für den akademischen Unterricht.* Tübingen: Stauffenburg.

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

Tsohatzidis, S. L. (ed.) (1994). *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives*. London: Routledge.

Ulkan, M. (1993). *Zur Klassifikation von Sprechakten. Eine grundlagen-theoretische Fallstudie*. Tübingen: Niemeyer.

10.2 HOMOGENITA, PODOBNOST A HIERARCHIE POSTAV

Hypotéza

Realizace mluvních aktů je u každé postavy dramatu jiná. Ověřte tuto hypotézu.

Postup

Hypotéza vychází z racionální úvahy. Pokud by každá postava hry realizovala shodné mluvní akty, chybělo by jakékoli napětí, mezi jednotlivými rolemi by nebyl žádný rozdíl. K ověření této hypotézy použijte následující dvě metody:

(1) Použijte své třídy mluvních aktů (srov. část 10.1, „Frekvenční distribuce mluvních aktů") a jejich frekvence a na základě ranků provádějte různé testy homogenity. Mnoho takových testů lze najít v každé učebnici pojednávající o neparametrické statistice. Porovnejte jednotlivé postavy mezi sebou. Vyvoďte nějaké závěry.

(2) Za použití chí-kvadrát testu homogenity porovnejte frekvence typů mluvních aktů všech dvojic postav. Dosáhli jste stejných výsledků jako u rankových testů? Bylo by možné vymezit svébytné třídy postav (samozřejmě pouze v případě, že jich bude minimálně deset) nebo jsou vzájemně provázány?

(3) Prostřednictvím míry významnosti u každé dvojice postav získané na základě chí-kvadrát testu sestavte graf daného jednání nebo celého

dramatu, tj. jednotlivé osoby spolu vzájemně propojte, pouze pokud se realizace jejich mluvních aktů významně neliší.

(4) Vyhodnoťte vlastnosti získaného grafu (Balakrishnan 1997, West 2001). Vyhodnoťte vlastnosti každé postavy. Proveďte detailnější rozbor struktury dané hry.

(5) Pomocí indikátoru pravděpodobnosti vyhodnoťte na základě frekvencí typů mluvních aktů podobnost divadelních postav. Vytvořte vážený graf podobností postav, přičemž indikátor podobnosti považujte za váhu hrany. Jelikož žádná z postav nestojí vůči jiné izolovaně, na základě součtu jejich podobností s dalšími postavami sestavte jejich hierarchii (centralitu).

Literatura

Balakrishnan, V. K. (1997). *Graph theory*. New York: McGraw-Hill.

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

10.3 VZDÁLENOSTI MEZI STEJNÝMI MLUVNÍMI AKTY

Hypotéza

Vzdálenosti mezi stejnými mluvními akty v dramatickém díle podléhají určité struktuře. Identifikujte tuto strukturu.

Postup

Jedním ze způsobů identifikace struktury je zkoumání (pozičních) vzdáleností mezi stejnými mluvními akty (tj. mluvními akty téhož druhu). Text je třeba transkribovat na sekvence symbolů mluvních aktů, případně anotovat takovými sekvencemi, postavy v tomto případě nejsou relevantní. K dosažení spolehlivých výsledků je nutné provést rozbor minimálně jednoho celého jednání

určitého dramatu. Jelikož jednotlivé postavy vykazují své vlastní postoje, řečové zvyklosti a komunikační strategie, je možné vyslovit předpoklad, že zde platí Skinnerova hypotéza, a sice že při krátké vzdálenosti narůstá pravděpodobnost výskytu téže jednotky (srov. Strauss et al. 2014, kap. „Fonetická agregace"). Z toho plyne, že krátkých vzdáleností je více než dlouhých. Vzdálenost je možné měřit počtem různých mluvních aktů mezi dvěma rovnocennými akty a přičtením hodnoty 1 nebo počtem „kroků" mezi dvěma sousedními identickými mluvními akty.

Podle této hypotézy nejsou vzdálenosti distribuovány rovnoměrně (se stejnou pravděpodobností), ale představují monotónně klesající sekvenci. Aproximujte tuto sekvenci prostřednictvím Zipf-Aleksejevovy funkce

$$y = ax^{-b-c\ln x},$$

kde x je vzdálenost (x = 1, 2, 3, ...), y je počet výskytů této vzdálenosti a a, b, c jsou parametry. Pokud výše uvedená funkce není vhodná, najděte nějakou adekvátnější.

Aplikujte tento postup na každou část textu zvlášť a pozorně sledujte vývoj parametrů od první k poslední části. Následně doplňte vzdálenosti a analyzujte hru jako celek.

Pokuste se v sekvenci vzdáleností odhalit nějaké další struktury.

Literatura

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.

Skinner, B. F. (1957). *Verbal behaviour.* Acton: Copley.

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

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

10.4 ŠKÁLOVÁNÍ MLUVNÍCH AKTŮ

Problém

Stanovte nějaký způsob škálování pro jednotlivé druhy mluvních aktů.

Postup

V určitém bodě vědeckého pokroku již kvalitativní klasifikace dále neumožňuje získávat hlubší vhled do mechanismů dané sféry zájmu. V tomto ohledu je třeba se řídit Galileovým výrokem „…měřit měřitelné, a neměřitelné měřitelným učinit". Vytvoření škály zde znamená zobrazit mluvní akty v nějaké konkrétní dimenzi. Může se jednat o různé dimenze: status mluvčího, postoj mluvčího k posluchači, emoce vyjadřované daným typem mluvního aktu, váha či intenzita mluvního aktu, např. obraty, jimiž se člověk *ptá, prosí, vyžaduje, naléhá* či *nařizuje*, mají různé parametry „naléhavosti" či „váhy" a současně vyjadřují určitý postoj.

Než se však pustíme do sestrojování škál, měli bychom si formulovat hypotézy, které je nutné či přinejmenším vhodné pomocí tohoto škálování testovat. Mohou nám to ukázat následující příklady: (a) čím je postava dramatu dominantnější, tím větší „váhu" mají její mluvní akty nebo (b) stupeň emocionality (mluvních aktů nějaké postavy) je funkcí protagonismu atd. Dominanci a protagonismus je samozřejmě nutné měřit nezávisle.

Tímto způsobem lze dospět k přesnějším kvantitativním konceptům. Nebyla by zde již vlastně nutná žádná klasifikace, měli bychom k dispozici měřitelné vlastnosti a analyzovaný text by bylo možné transkribovat jako sekvenci dvojic atribut-hodnota (nebo, ve zkráceně formě v případě jediné analyzované dimenze, jednoduše jako numerické hodnoty).

Každý badatel zabývající se studiem mluvních aktů má intuitivní představu o „váze" určité fráze. Pokud se budeme snažit převést tuto intuici na čísla, je možné navrhnout nějakou škálu. Jako analogie zde může posloužit zdvořilost (část 9.2, „Zdvořilost"), jejíž míru může intuitivně odhadnout každý rodilý mluvčí.

Literatura
žádná

10.5 DISTRIBUCE ŠKÁLOVANÝCH HODNOT MLUVNÍCH AKTŮ

Hypotéza
Distribuce hodnot mluvních aktů nejsou v jednotlivých jednáních dramatu homogenní. Ověřte tuto hypotézu.

Postup
Přepíšou-li se jednotlivé mluvní akty dramatu jako hodnoty na škále vzniklé na základě zadání v části 10.4, „Škálování mluvních aktů", dostaneme sekvenci čísel. Máme-li frekvenci výskytu jednotlivých stupňů, můžeme současně získat frekvenční distribuci dané vlastnosti. Jelikož se však drama vyznačuje jistým dynamickým průběhem – počáteční konflikt postupně spěje k svému vyvrcholení, které pak přechází v katarzi – každé jednání bude obsahovat mluvní akty různého stupně. Je možné předpokládat, že všechny distribuce se budou řídit týmiž principy, ale jejich formy (parametry) budou natolik rozdílné, že budou nutně vykazovat výraznou heterogenitu. Heterogenitu lze testovat za použití chí-kvadrát testu a dynamiku dramatu lze například popsat jako sekvenci průměrů jednotlivých stupňů.

Další hypotézy o vzniku distribucí bude možné vyslovit, jakmile budou generována první data.

Literatura
žádná

10.6 MOTIVY VÁHY

Problém
Postupná váha hodnot intenzity mluvních aktů (srov. předchozí zadání) vytváří sekvence v analogii ke Köhlerově motivům. Sekvence hodnot váhy jsou ve skutečnosti jen další formou těchto motivů. Všechny výzkumné postupy a metody, o nichž se hovoří v příslušné literatuře, lze tudíž rovněž aplikovat na hodnoty vlastností mluvních aktů.

Postup
Zvolte si jeden z možných rámců mluvních aktů, tj. celé texty, jednotlivá jednání nebo kapitoly, mluvní akty konkrétních jednotlivců apod. Začněte u sekvence hodnot, jež je výsledkem jednoho z předchozích řešených problémů, a podle definice Köhlerových motivů vytvořte příslušné jednotky (viz literatura): motiv začíná na začátku daného rámce (tj. textu nebo jednání) či tam, kde končí předchozí motiv. Aktuální motiv končí ve chvíli, kdy je další hodnota menší než hodnota současná. Takže například sekvence hodnot

2-3-3-5-3-4-4-2-1-3-5

je segmentována na motivy

2-3-3-5, 3-4-4, 2 a 1-3-5.

Tyto motivy jsou jednotkami, které lze zkoumat z hlediska jejich frekvencí (tj. jak často se motiv 3-4-4 vyskytuje v daném rámci), délky (např. motiv 2-3-3-5 má délku 4) atd. Analyzovat tak můžete rankovou distribuci motivů (můžete očekávat Waringovo nebo Zipf-Mandelbrotovo rozdělení), délkovou distribuci (hyper-Pascalovo nebo hyper-Poissonovo rozdělení) atd.

Tyto studie lze navíc provádět také ve větším měřítku. Můžete například zkoumat délku délkových motivů, tj. uvažovat sekvenci hodnot délky motivů prvního řádu jako novou úroveň analýzy a vytvořit na této úrovni nové motivy odpovídající výše uvedené definici. Tento postup je možné opakovat, dokud na poslední úrovni nezůstane jen velmi malé množství motivů.

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. (2008). Word length in text. A study in the syntagmatic dimension. In: Mislovičová, S. (ed.), *Jazyk a jazykoveda v pohybe.* Bratislava: VEDA, 416–421.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, B., Schmidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications. Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg.* Berlin, Heidelberg: Springer, 637–646.

10.7 DRAMA JAKO ČASOVÁ ŘADA MLUVNÍCH AKTŮ

Problém

Najděte diferenční rovnici nejnižšího řádu simulující sekvenci škálovaných mluvních aktů v nějakém dramatu.

Postup

Jsou-li mluvní akty škálovány, pak je text možné zpodobnit jako sekvenci čísel. Jelikož však divadelní hra spočívá v ustavičném střetu mluvních aktů a jednotlivé postavy se ve svých promluvách střídají, mohou se s tím proměňovat také stupně mluvních aktů. Typ výsledného střídání (kolísání) bude záviset na povaze hry. Toto kolísání lze každopádně dobře postihnout nejnižším stupněm diferenční rovnice. Příslušný software takové výsledky vygeneruje mechanicky.

Na základě analýzy několika her najděte vzájemnou souvislost mezi určitým druhem hry a řádem diferenční rovnice. Pokuste se o její interpretaci.

Literatura

žádná

10.8 NĚKTERÉ VLASTNOSTI SEKVENCÍ MLUVNÍCH AKTŮ

Problém

Najděte a vypočítejte některé další vlastnosti sekvence hodnot mluvních aktů.

Postup

Zaměřte se na každé jednání konkrétního dramatu zvlášť. Jednotlivé mluvní akty nahraďte stupněm na nějaké škále. Na základě získané sekvence čísel vypočtěte Hurstův exponent, Ljapunovův koeficient a Minkowského klobásu, srov. Strauss et al. (2014, kap. „Hurstův exponent", „Ljapunovův koeficient" a „Minkowského klobása"). Zamyslete se nad změnami těchto kvantit v průběhu dramatického díla. Mění se, nebo jsou konstantní? Je možné na základě těchto čísel vysvětlit chování mluvních aktů nebo pozadí dané hry? Jsou tyto koeficienty v korelaci s jinými vlastnostmi divadelních her?

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). *Lectures on text theory.* Prague: Oriental Institute.

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.

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

Strauss, U., Fan, F., Altmann, G. (2014). *Kvantitativní lingvistika. Vybrané problémy 1*. Olomouc: Univerzita Palackého v Olomouci.

10.9 DRAMA A KOMEDIE

Problém

Porovnejte drama a komedii z hlediska všech aspektů mluvních aktů.

Postup

Proveďte srovnání minimálně jednoho dramatu a jedné komedie prostřednictvím veškerých kvantitativních dat získaných v rámci řešení zadání v částech 10.1 až 10.8. Které aspekty vykazují největší rozdíly? Proveďte interpretaci rozdílů i shodných rysů a stanovte pro divadelní hry určitá pravidla.

Literatura

žádná

10.10 VÝVOJ DRAMATU

Problém

Vzhledem k rozdílům mezi antickými řeckými dramaty a moderními divadelními hrami dochází nutně k postupné proměně. Pokuste se ji prostřednictvím mluvních aktů postihnout.

Postup

Proveďte analýzu dramat z různých epoch v jednom či několika různých jazycích. Nejprve si sestavte soubor všech mluvních aktů, kvantitativně je zpracujte, vypočtěte charakteristické rysy (srov. části 10.1 až 10.9) a demonstrujte jejich vývoj, kulturní rozdíly apod.

Literatura

žádná

10.11 HREBY MLUVNÍCH AKTŮ

Problém

Je možné vytvořit na základě mluvních aktů hreby?

Postup

Přepište text nějakého dramatu ve formě mluvních aktů. Následně uvažujte, že všechny věty určité postavy, které obsahují tentýž mluvní akt, náleží k témuž hrebu. Každá věta může náležet i k několika různým hrebům. Stanovte inventář hrebů a jeho rozsah u každé postavy ve smyslu počtu vět v nich obsažených. Ukažte, že postavy dramatu se v tomto ohledu liší. Definice hrebu viz Strauss et al. (2014, kap. „Hreby“).

Určete frekvenční distribuci velikosti hrebu a odvoďte rozdělení pravděpodobnosti nebo alespoň induktivně vyvoďte distribuci, která ji vystihuje.

Liší se v tomto ohledu dramata od komedií?

Je možné určit jednotky sestávající ze sekvencí mluvních aktů? Existují sekvence, které jsou pro určité postavy charakteristické?

Vytvořte v tomto směru „teorii" mluvních aktů.

Literatura

Srov. část 10.1

10.12 SMĚREM K TEORII MLUVNÍCH AKTŮ

Problém

Je nauka o mluvních aktech teorií nebo pouze utříděným souborem definovaných pojmů? Předložte argumenty pro jedno z těchto pojetí.

Postup

Pokud je nauka o mluvních aktech něco více než jen dobře utříděný soubor pojmů, jehož jedna část je dobře definovaná, zatímco druhá ne zcela jasně, předložte argumenty pro její teoretický status. Prostudujte si dostupnou literaturu a shromážděte hypotézy týkající se určitých vztahů, závislostí, vývoje či systémového statutu mluvních aktů. Pokud to bude nutné, podpořte své argumenty psychologicky nebo sociologicky. Vyjádřete tyto hypotézy formálně, odvoďte je z některých předběžných axiomů (domněnek) nebo navrhněte alespoň nějaký vzorec. Každopádně však prokažte, že danou hypotézu lze ověřit, a poskytněte alespoň náznak možnosti nějakého objektivního způsobu ověření.

Pokud takovou možnost nenajdete v literatuře, vytvořte nějaké hypotézy.

Literatura

Srov. část 10.1

10.13 DÉLKA DIALOGOVÝCH PŘÍSPĚVKŮ

Problém

Určete distribuci délek příspěvků v dialozích/polylozích.

Postup

Určete délky jednotlivých příspěvků v rámci dialogů v divadelních hrách, filmových scénářích a spontánních promluvách zachycených v korpusech mluveného jazyka. Přiřaďte jednotlivé příspěvky jejich původcům a vypočítejte frekvenční distribuce délek příspěvků (ve smyslu počtu vět) (a) pro text jako celek a (b) pro jednotlivé aktéry. Délkové hodnoty budete muset sdružit do intervalů, např. 1–5, 6–10 apod., abyste zajistili dostatek dat ve všech třídách.

(a) Které rozdělení pravděpodobnosti bude podle vašeho očekávání odpovídat shromážděným datům? Proveďte aplikaci zvoleného rozdělení na daná data a svou hypotézu ověřte. Výslednou distribuci zdůvodněte lingvistickými argumenty.
(b) Lze u jednotlivých aktérů pozorovat výrazné rozdíly v distribuci (i) v rámci jednoho textu, (ii) mezi různými texty a typy textů?
(c) Jsou distribuce a jejich parametry v nějakém vztahu k (i) počtu příspěvků jednotlivých aktérů, (ii) počtu aktérů polylogu, (iii) (sociálnímu) statutu osob účastnících se dialogu/polylogu?

Literatura

žádná

10.14 DISKURZNÍ FREKVENCE (1)

Problém

Určete rankovou distribuci gramatických kategorií v diskurzu.

Postup

V závislosti na jazyku nebo jazycích, které se chystáte zkoumat, si zvolte jednu nebo více gramatických kategorií: pád, číslo, rod, osobu, čas, vid, diatezi, určitost, atd. Sestavte si soubor z textů jednoho druhu, např. román, zpráva, novinový text, interview, dialog atd. nebo použijte vhodný (sub)korpus. Spočítejte různé gramatické rysy podle zvolených kategorií a zpracujte si frekvenční tabulku (distribuci). Seřaďte jednotlivé formy podle jejich frekvence.

Je možné aplikovat teoretické rozdělení pravděpodobnosti na získaná data?

Tip: Pokud jsou vaše texty příliš dlouhé, analyzujte jednotlivé texty postupně. V opačném případě mohou vyvstat dva metodologické problémy: (1) nehomogenita dat a (2) přílišná velikost datových souborů, která může způsobit nefunkčnost chí-kvadrát testu.

Pokud se vám nepodaří aplikovat na data rozdělení pravděpodobnosti, použijte prostou spojitou funkci nebo řadu, tj. s vynecháním normalizace. Pokud ani nadále nebude možné vaše data použít k modelování, najděte extrémní hodnoty a prostřednictvím lingvistických argumentů vysvětlete jejich odchýlení od obecného trendu.

Literatura

Altmann, G. (1992). Das Problem der Datenhomogenität. *Glottometrika 13,* 105–120.

Myhill, J. (2005). Quantitative methods of discourse analysis. In: Köhler, R., Altmann, G., Piotrovskij, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook.* Berlin, New York: de Gruyter, 471–798.

10.15 DISKURZNÍ FREKVENCE (2)

Hypotéza

Bloková frekvenční distribuce gramatických kategorií v textech podléhá zákonu Frumkinové. Ověřte tuto hypotézu.

Postup

V analogii k blokové distribuci funkčních slov (popisované Frumkinovou [1962] a dalšími – viz literatura) a syntaktických konstrukcí/funkcí (srov. Köhler 2001) určete počet textových bloků (zkuste bloky o velikosti 10, 30, 50 a 100 slov) s výskytem 0, 1, 2,… zvolených gramatických kategorií (např. plurál, duál, genitiv, futurum atd.). Aplikujte na daná data negativně hypergeometrické rozdělení („Frumkinové zákon"). Pozorujte závislost hodnot parametrů na délce bloků, počtu bloků a typu kategorie. Negativně hypergeometrické rozdělení je definováno jako

$$P(X=x)=\frac{\binom{M+x-1}{x}\binom{K-M+n-x-1}{n-x}}{\binom{K+n-1}{n}}\ ,x=0,1,\ldots,n\ ,$$

kde K, M a n jsou parametry.

Literatura

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

Altmann, G., Burdinski, V. (1982). Towards a law of word repetitions in text-blocks. *Glottometrika 4,* 146–167.

Bektaev, K. B., Luk´janenkov, K. F. (1971). O zakonach raspredelenija edinic pis´mennoj reči. In: Piotrowski, R. H. (ed.), *Statistika reči i avtomatičeskij analiz teksta.* Leningrad: Nauka, 47–112.

Brainerd, B. (1972). Article use as an indirect indicator of style among English-language authors. In: Jäger, S. (ed.), *Linguistik und Statistik.* Braunschweig: Vieweg, 11–32.

Francis, I. S. (1966). An exposition of a statistical approach to Federalist dispute. In: Leed, J. (ed.), *The computer and literary style.* Kent, Ohio: Kent State University Press, 38–78.

Frumkina, R. M. (1962). O zakonach raspredelenija slov i klassov slov. In: Mološnaja, T. N. (ed.), *Strukturno-tipologičekie issledovanija 124.33*. Moskva: Akademija Nauk SSSR.

Köhler, R. (2001). The distribution of some syntactic construction types in text blocks. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček.* Trier: Wissenschaftlicher Verlag, 136–148.

Maškina, L. E. (1968). *O statističeskich metodach issledovanija leksiko-gramatičeskoj distribucii*. Minsk, Diss.

Mosteller, F., Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist.* Reading, Mass: Addison-Wesley.

Paškovskij, V. E., Srebrjanskaja, I. I. (1971). Statističeskie ocenki pis´mennoj reči bol'nych šizofreniej. In: *Inženernaja lingvistika*. Leningrad: Nauka.

Piotrowski, R. G. (1984). *Text, Computer, Mensch*. Bochum: Brockmeyer.

Wimmer, G., Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

10.16 DISKURZNÍ FREKVENCE (3)

Hypotéza

Podíl nezměněných slov, derivativ, kompozit, odvozených kompozit apod. v textu závisí na typu a stáří textu. Ověřte tuto hypotézu.

Postup

Sestavte si soubor textů různého typu a autorství. Každý typ textu a autora by měl být zastoupen nějakým číslem. Následně u každého slovotvorného typu určete počet slov, která jej reprezentují, a vypočítejte jejich podíly v jednotlivých textech. Na základě provedení statistického testu zjistěte, zda jsou existující rozdíly signifikantní a zda mohou být pro jednotlivé typy textů anebo autory charakteristické. Identifikujte takové způsoby tvoření slov, které jsou pro daný text nejspecifičtější, tj. jejichž podíly se významně liší od podílů v jiných textech nebo skupinách textů. Sestavte rankové pořadí způsobů tvoření slov podle jejich diskriminační síly. Proveďte interpretaci výsledků a lingvisticky nebo psychologicky je zdůvodněte.

Literatura

učebnice statistiky

10.17 RÉTORICKÁ STRUKTURA (1)

Problém

Proveďte analýzu několika textů z hlediska kategorií „teorie rétorické struktury" (TRS) a určete frekvenční distribuci těchto kategorií.

Postup

Několik textů označte na základě analýzy TRS tagy. Zjistěte počet výskytů různých tagů bez ohledu na jejich pozici v dané struktuře.

Je možné na tato data aplikovat teoretické rozdělení pravděpodobnosti?

Jakou distribuci či jaký typ distribuce očekáváte? Jak by se případně daly výsledné distribuce interpretovat či dokonce vysvětlit?

Literatura

Mann, W. C., Thompson, S. A. (1988). *Rhetorical structure theory: toward a functional theory of text organization*. Text 8(3), 243–281.

RST: Rhetorical Structure Theory. [online]. Dostupné z: http://www.sfu.ca/rst/01intro/intro.html (cit. 23. září 2009).

10.18 RÉTORICKÁ STRUKTURA (2)

Problém

Je distribuce tagů TRS závislá na pozici těchto tagů v rámci dané struktury?

Postup

Analyzujte několik textů na základě TRS. Zjistěte počet výskytů různých tagů s ohledem na jejich pozici v dané struktuře: vypočtěte frekvence zvlášť pro

(a) jednotlivé roviny hierarchie,

(b) pozici v textu danou počtem tagů od začátku textu,

(c) pozici v dílčí struktuře danou počtem tagů od začátku dílčí struktury.

Je možné na tato data aplikovat teoretické rozdělení pravděpodobnosti?

Jakou distribuci či jaký typ distribuce očekáváte? Jak by se případně daly výsledné distribuce interpretovat či dokonce vysvětlit?

Literatura

Mann, W. C., Thompson, S. A. (1988). *Rhetorical structure sheory: toward a functional theory of text organization*. Text 8(3), 243–281.

RST: Rhetorical Structure Theory. [online]. Dostupné z: http://www.sfu.ca/rst/01intro/intro.html (cit. 23. září 2009).

10.19 RÉTORICKÁ STRUKTURA (3)

Problém

Definujte některé vlastnosti jednotek TRS.

Postup

Jako každé jednotce, i jednotkám TRS je možné přidělit množství zajímavých vlastností. Definujte komplexitu a další vlastnosti v analogii k vlastnostem syntaktických struktur, jak je definuje Köhler (1999).

Literatura

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1),* 46–57.

RST: Rhetorical Structure Theory. [online]. Dostupné z: http://www.sfu.ca/rst/01intro/intro.html (cit. 23. září 2009).

10.20 RÉTORICKÁ STRUKTURA (4)

Problém

Formulujte hypotézy o vzájemných vztazích mezi vlastnostmi jednotek TRS a ověřte je.

Postup

Postulujte vzájemné vztahy (závislosti) mezi úrovněmi uzlů v rámci struktury TRS, pozicemi v textu nebo v dané dílčí struktuře, frekvencemi nebo komplexitou jednotek TRS a dalšími vlastnostmi.

Ověřte tyto hypotézy na datech získaných na základě řešení zadání z části 10.18, „Rétorická struktura (2)". Jako vodítko lze použít analogickou studii Köhlera (1999) zabývající se touto problematikou v syntaktické rovině.

Literatura

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1)*, 46–57.

RST: Rhetorical Structure Theory. [online]. Dostupné z: http://www.sfu.ca/rst/01intro/intro.html (cit. 23. září 2009).

Rejstřík jmenný

C

D

E

F

G

H

I

J

K

L

M

N

O

P

R

S

T

U

V

W

Y

Z

KATALOGIZACE V KNIZE - NÁRODNÍ KNIHOVNA ČR

Köhler, Reinhard

Kvantitativní lingvistika : vybrané problémy 2 / Reinhard Köhler, Gabriel Altmann ; [překlad Miroslav Kubát, Radek Čech]. -- 1. vyd. -- Olomouc : Univerzita Palackého v Olomouci, 2014. -- 207 s. -- (Qfwfq ; sv. 32)
Název originálu: Problems in quantitative linguistics 2
Přeloženo z angličtiny
ISBN 978-80-244-4324-9

*** 81'324**
- kvantitativní lingvistika
- monografie

81 - Lingvistika. Jazyky [11]

Kvantitativní lingvistika. Vybrané problémy 2
Reinhard Köhler, Gabriel Altmann

32. svazek Edice Qfwfq

Výkonný redaktor: Jiří Špička
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, 208 stran
č. z. 2014/814

ISBN 978-80-244-4324-9