Výzkum variability v korpusech češtiny

Václav Cvrček, Pavel Vondřička
Ústav Českého národního korpusu, Filozofická fakulta,
Univerzita Karlova v Praze

Abstract

This paper presents the project of a new tool for exploring the variability of language in corpora of Czech, and describes its design. SyD makes this type of corpus exploration even easier for ordinary language users. The straightforward Google-like opening form of this web-based application allows users to enter two or more variants. These variants, i.e. corpus queries, can be in the form of simple word forms, phrases and CQL queries (including lemmas and morphological tags). The application then searches for each variant in the Czech National Corpus. SyD has two main components: the synchronic and the diachronic (hence SyD). The synchronic part searches for variants in corpora of contemporary spoken and written texts (these are SYN2010, a 100 million word representative corpus of written Czech, and Oral2006 plus Oral2008, corpora of informal spoken Czech, 1 million words each). It calculates the overall absolute and relative frequency of each variant plus its specific frequencies for various subcorpora – e.g. how many times each variant occurs in specific text types, genres, or how often it occurs in the speech of different groups of speakers (according to sex, age, education and region) etc. The diachronic part is applied to a special corpus called Diakon, which was created essentially to cover the history of the Czech language from the 14th century until the present day. At present this corpus contains almost 140 million words in more than 3,800 texts. The final chronological visualisation shows change in the proportions of the variants over time.

1. Úvod

Jazyková variabilita představuje dlouhodobě jedno z nejzajímavějších lingvistických témat. Jazykovědce láká především z toho důvodu, že je živnou půdou spontánního jazykového vývoje, který se odehrává především prostřednictvím variant a oscilace mezi nimi (Cvrček 2008: 34). Zároveň je jazyková variabilita základem stylové a žánrové rozrůzněnosti textů; předpoklady pro vytváření obsahově shodných textů s různou stylovou platností jsou tedy už v langue (přistupujeme k variabilitě s tím, že nejde o chybu realizace, ale součást systému).

Vedle toho jsou varianty už tradičně v našich zeměpisných šířkách předmětem jazykové regulace. Ta se v drtivé většině případů pokouší o snížení variability s odůvodněním, že se tím zvyšuje prediktabilnost projevů, a tedy i zlepšuje jejich percepce.[3] Je přitom potřeba připomenout, že variantnost se vyskytuje ve všech oblastech jazyka, tedy i v těch, které oficiální jazykové regulaci nepodléhají. Nezanedbatelná míra variability tak projde testem účelnosti a jazykové ekonomie, které jsou jediným měřítkem funkčnosti jazykových jevů ve varietách vyvíjejících se spontánně a bez vážnějších zásahů lingvistů.

Výzkum jazykové variability s příchodem korpusů dostává nový impulz. Vedle přesnějšího určení poměru jednotlivých variant můžeme ve velkých a stylově rozmanitých korpusech identifikovat i rozsah variability. Důraz je tak dnes kladen nejen na popis jednotlivých variant spolu s jejich vzájemnými poměry (viz např. Mluvnice současné češtiny, Cvrček et al. 2010), ale i na jejich funkční odlišení.

Variabilita se přitom nemusí nutně odehrávat – jak by mohla naznačovat výše uvedená poznámka o současné české jazykové regulaci – pouze v rovině morfologie. Korpusy umožňují zkoumat variabilitu lexikální i syntagmatickou (varianty slovosledné). Vedle toho se objevují i nové roviny jazyka, na nichž dříve variabilita zkoumána nebyla a přitom ji korpusy vyjevují (např. frazeologie).

Smyslem nástroje, který hodláme představit, je ulehčit laickému i poučenému uživateli přístup k výzkumu jazykových variant v korpusech ČNK, aby nemusel věnovat téměř žádné úsilí technické stránce hledání a mohl se plně soustředit na analýzu a interpretaci nalezených dat. Zároveň by představovaný nástroj měl být doplňkem k tradičním kodifikačním příručkám (ať už papírovým nebo internetovým), které nabízejí tu více tu méně černobílou odpověď ve stylu dobře-špatně, nebo – v horším případě – se o některých variantách z důvodu jejich „nesprávnosti" nezmiňují vůbec.

2. SyD: Návrh aplikace

SyD je webová aplikace, která slouží k všestrannému výzkumu variant v korpusech ČNK. Její návrh respektuje základní badatelská východiska a odlišnosti v typech textů. Uživatel zadává dvě nebo víc variant a průzkum se provádí v následujících oblastech:

Diachronie
    1. celá historie psané češtiny
    2. jednotlivá období
Synchronie
    1. Psaný jazyk
    1.1. Textové registry (textové makrotypy)
    1.2. Textové typy
    1.3. Žánry
    1.4. Média a překladovost/originálnost textu

---

[3] Např. *„Dodržování relativně pevného standardu usnadňuje percepci mluveného projevu, protože se zvyšuje jeho prediktabilita...* (...) *Rozšiřování tolerance, např. zaváděním množství dublet, znamená tuto výhodu standardu oslabit či anulovat."* (Palková 1993: 77) Otázkou zůstává, zda je tato prediktabilita opravdu specifickou vlastností spisovného standardu a není prostě charakteristická pro jakýkoli soubor pravidelně užívaných tvarů (úzus), nikoli nutně spisovných (tj. vynucených standardizací). (Cvrček 2006: 96).

2. Mluvený jazyk
    2.1. Pohlaví a věk mluvčích
    2.2. Vzdělání mluvčích
    2.3. Regionální příslušnost mluvčích

Celý projekt je postaven na spolupráci korpusového serveru Manatee, jehož autorem je P. Rychlý, skriptů v jazyce Perl, které zajišťují konstrukci dotazů, a uživatelského rozhraní napsaného v jazyce PHP a HTML stránek rozšířených o interaktivitu pomocí JavaScriptu. Grafická část je obstarávána knihovnami Google Charts API a jQuery.

Testovací provoz. Děkujeme za vaše připomínky.

SyD je určen pro všestranný průzkum variant jak v současném jazyce - **synchronní část**, tak v průběhu jeho vývoje - **diachronní část**. K analýze využívá data mluvených a psaných korpusů projektu Českého národního korpusu. Stačí zadat dvě varianty (nebo víc) jednoho jevu, které si vzájemně konkurují (např. *ačkoli* × *ačkoliv*, *už* × *již*, *mimoto* × *mimo to*, *sama sebe* × *sebe sama*, *stále* × *pořád* × *furt* apod.).

Vyhodnocení složitejších dotazů, jejichž podoba může využít všech možností dotazovacího jazyka CQL, může trvat i několik desítek sekund.

Poslední položené dotazy:

adorovat
zaprvé × za prvé
skopávat × zkopávat
skopávat × zkopávat
podpaží × podpaždí

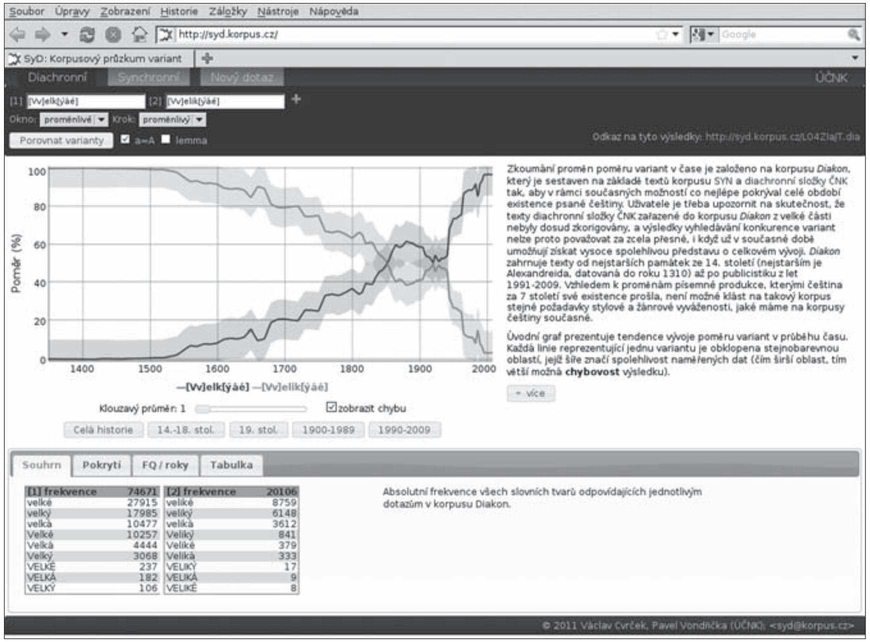
Obr. 1: Úvodní obrazovka aplikace SyD, zadávání dotazů.

2.1 Diachronní část

Zkoumání proměn poměru variant v čase je založeno na korpusu Diakon, který je sestaven na základě textů korpusu SYN a diachronní složky ČNK tak, aby v rámci současných možností co nejlépe pokrýval celé období existence psané češtiny. Texty zařazené do korpusu Diakon z velké části nebyly dosud zkorigovány, a výsledky vyhledávání konkurence variant nelze proto považovat za zcela přesné, i když už v současné době umožňují získat vysoce spolehlivou představu o celkovém vývoji. Korpus není lemmatizován ani morfologicky označkován (což samozřejmě limituje možnosti dotazování).

Diakon zahrnuje texty od nejstarších památek ze 14. století (nejstarším je Alexandreida, datovaná do roku 1310) až po publicistiku z let 1991-2009 (těchto posledních 20 let je zastoupeno publicistikou největších deníků LN, MfD, Právo a HN v objemu 5 milionů textových slov za rok). Vzhledem k proměnám písemné produkce, kterými čeština za 7 století své existence prošla, není možné klást na takový korpus stejné požadavky stylové a žánrové vyváženosti, jaké máme na korpusy češtiny současné.

Uživateli se po zadání dotazu v diachronní složce objeví úvodní graf, který zobrazuje tendence vývoje poměru variant v průběhu času. Každá linie reprezentující jednu variantu je obklopena stejnobarevnou oblastí, jejíž šíře značí spolehlivost naměřených dat (čím širší oblast, tím větší možná chybovost výsledku).



[1] [Vv]elk[ýáé]
[2] [Vv]elik[ýáé]

Odkaz na tyto výsledky: http://syd.korpus.cz/L04ZlajT.dia

Zkoumání proměn poměru variant v čase je založeno na korpusu *Diakon*, který je sestaven na základě textů korpusu SYN a diachronní složky ČNK tak, aby v rámci současných možností co nejlépe pokrýval celé období existence psané češtiny. Uživatele je třeba upozornit na skutečnost, že texty diachronní složky ČNK zařazené do korpusu *Diakon* z velké části nebyly dosud zkorigovány, a výsledky vyhledávání konkurence variant nelze proto považovat za zcela přesné, i když už v současné době umožňují získat vysoce spolehlivou představu o celkovém vývoji. *Diakon* zahrnuje texty od nejstarších památek ze 14. století (nejstarším je Alexandreida, datovaná do roku 1310) až po publicistiku z let 1991-2009. Vzhledem k proměnám písemné produkce, kterými čeština za 7 století své existence prošla, není možné klást na takový korpus stejné požadavky stylové a žánrové vyváženosti, jaké máme na korpusy češtiny současné.

Úvodní graf prezentuje tendence vývoje poměru variant v průběhu času. Každá linie reprezentující jednu variantu je obklopena stejnobarevnou oblastí, jejíž šíře značí spolehlivost naměřených dat (čím širší oblast, tím větší možná **chybovost** výsledku).

[1] frekvence	74671	[2] frekvence	20106
velké	27915	veliké	8759
velký	17985	veliký	6148
velká	10477	veliká	3612
Velké	10257	Veliký	841
Velká	4444	Veliké	379
Velký	3068	Veliká	333
VELKÉ	237	VELIKÝ	17
VELKÁ	182	VELIKÁ	9
VELKÝ	106	VELIKÉ	8

Absolutní frekvence všech slovních tvarů odpovídajících jednotlivým dotazům v korpusu Diakon.

Obr. 2: Diachronní analýza variant *velký/velká/velké* vs. *veliký/veliká/veliké*.

Výpočet chybovosti se odvíjel od následujících premis: 1) nejpřesnější výsledky v daném období (časovém okně) je analýza schopna poskytnout o nejfrekventovanějším typu (slovu). Naopak, 2) největší chybu by měly vykazovat dotazy, kde každá z variant má pouze jeden nebo žádný výskyt. 3) S přibývajícími doklady rost naše jistota (a klesá chybovost) logaritmicky, nikoli lineárně.[4] Jelikož by vzhledem k variabilitě možných nastavení časového okna a kroku (viz níže) bylo výpočetně velmi náročné zjišťovat frekvenci nejčastějšího typu pro každé období v reálném čase během zpracování dotazu, používá se pro výpočet chyby namísto toho odhad frekvence nejfrekventovanějšího typu, který se zhruba rovná 1/13 celkové velikosti (sub)korpusu. Výsledný vzorec pro výpočet chyby tak je:

chyba = ln [(*Počet dotazů × Velikost subkorpusu/13*) / *Souhrnná frekvence všech variant*]

V případech, kdy nejsou k dispozici žádné doklady ani u jedné z variant, je chyba nastavena na hodnotu 100 % vydělenou počtem variant, v případě, že souhrnná frekvence všech variant je větší než počet dotazů × velikost subkorpusu/13, je hodnota chyby nastavena na 0 %.

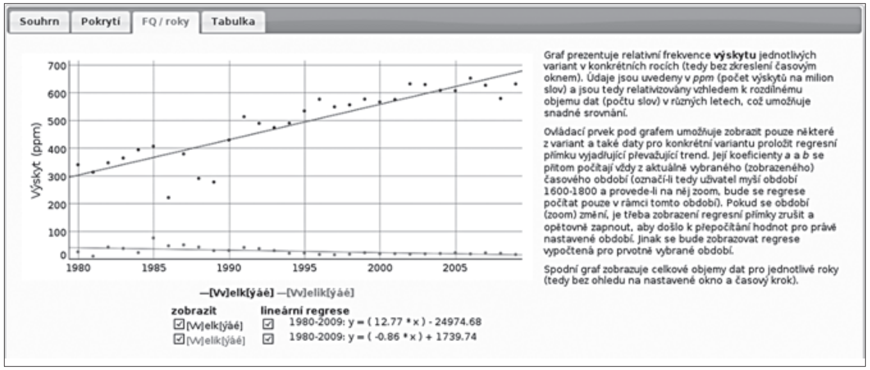
Hodnoty v úvodním grafu jsou výsledkem procházení korpusu v časových oknech po časových krocích. Např. hodnota vynesená pro rok 1550 (není-li požadováno jinak) je ve skutečnosti výsledkem zkoumání subkorpusu textů pocházejících z období 1500-1600. Velikost okna je možné nastavit v záhlaví stránky – ve výchozím nastavení je proměnlivá, což znamená, že se okno postupem od minulosti do současnosti mění (v období 1350-1800 je okno 100 let, od roku 1810 do roku 1900 je okno 50 let, v letech 1905-1983 je okno 20 let a konečně od roku 1986 je okno 3 roky). Zároveň je možné nastavit časový krok, tedy úsek, o který se okno vždy posouvá při procházení korpusu směrem od minulosti k současnosti. Ve výchozím nastavení je krok rovněž proměnlivý, a to takto: v nejstarším období 1350-1800 se okno posouvá v desetiletých krocích, od roku 1810 do roku 1900 je krok 5 let, ve 20. století je krok 3 roky a konečně od roku 1986 je krok jednoletý.

V grafu je možné podrobněji prohlížet výsek časové osy podle vlastního výběru, případně se posouvat v čase oběma směry.

Pro získání pohledu abstrahujícího od jednotlivých lokálních odchylek je možné použít vyhlazování pomocí klouzavých průměrů (ovládací prvek se nachází pod grafem). Míra vyhlazení (zobrazí se průměry z 1-5 hodnot) ovlivňuje jak celkovou podobu výsledku, tak i jejich chybovost.

Všechny dotazy a jejich výsledky (v synchronní i diachronní sekci) se archivují a jsou opakovatelně vyvolatelné pomocí odkazu na výsledky, který je uveden v záhlaví stránky a který může sloužit i pro citační účely.

Vedle informací o absolutních frekvencích všech slovních tvarů odpovídajících jednotlivým dotazům v korpusu Diakon, pokrytí dotazovaného problému daty a celkové velikosti subkorpusů pro jednotlivá období, nabízí diachronní část projektu SyD i přesnější představu o jednotlivých výskytech jevů v historii: v sekci FQ/roky se nacházejí dva grafy, které prezentují poměry výskytů variant v konkrétních rocích (tedy bez zkreslení časovým oknem). Údaje jsou uvedeny v ppm (počet výskytů na milion slov) a jsou tedy relativizovány vzhledem k rozdílnému objemu dat (počtu slov) v různých letech, což umožňuje snadné srovnání.



Graf prezentuje relativní frekvence **výskytu** jednotlivých variant v konkrétních rocích (tedy bez zkreslení časovým oknem). Údaje jsou uvedeny v *ppm* (počet výskytů na milion slov) a jsou tedy relativizovány vzhledem k rozdílnému objemu dat (počtu slov) v různých letech, což umožňuje snadné srovnání.

Ovládací prvek pod grafem umožňuje zobrazit pouze některé z variant a také daty pro konkrétní variantu proložit regresní přímku vyjadřující převažující trend. Její koeficienty *a* a *b* se přitom počítají vždy z aktuálně vybraného (zobrazeného) časového období (označí-li tedy uživatel myší období 1600-1800 a provede-li na něj zoom, bude se regrese počítat pouze v rámci tomto období). Pokud se období (zoom) změní, je třeba zobrazení regresní přímky zrušit a opětovně zapnout, aby došlo k přepočítání hodnot pro právě nastavené období. Jinak se bude zobrazovat regrese vypočtená pro prvotně vybrané období.

Spodní graf zobrazuje celkové objemy dat pro jednotlivé roky (tedy bez ohledu na nastavené okno a časový krok).

zobrazit	lineární regrese	
☑ [Vv]elk[ýáé]	☑	1980-2009: y = (12.77 * x) - 24974.68
☑ [Vv]elik[ýáé]	☑	1980-2009: y = (-0.86 * x) + 1739.74

Obr 3: Výsledky pro jednotlivé roky (bez zkreslení časovým oknem) pro období 1980-2009.

Ovládací prvek pod grafem umožňuje zobrazit pouze některé z variant, a také daty pro konkrétní variantu proložit regresní přímku vyjadřující převažující trend. Její koeficienty *a* a *b* se přitom počítají vždy z aktuálně vybraného (zobrazeného) časového období (označí-li tedy uživatel myší období 1980-2009 a provede-li jeho zvětšení, bude se regrese počítat pouze v rámci těchto let). Druhý graf v této sekci zobrazuje celkové objemy dat pro jednotlivé roky (tedy bez ohledu na nastavené okno a časový krok).

Všechny údaje prezentované grafy v diachronní části (s výjimkou sekce FQ/roky, viz výše) shrnuje samostatná tabulka. Pro každé dílčí časové období (okno) uvádí celkový počet výskytů jednotlivých variant a jejich vzájemný poměr (v procentech). Ve sloupci chyba měření je uvedeno číslo, které je v grafu prezentováno šířkou barevného pásma okolo každé z čar. Souhrnná frekvence je součtem absolutních frekvencí všech variant a její relativizovaná hodnota v ppm (počet výskytů na milion slov) je uvedena v následujícím sloupci. Poslední sloupec uvádí celkový objem dat, který je v daném časovém okně k dispozici. Pokud je nastaven proměnlivý časový krok i okno, je tabulka rozdělena na několik částí v místech, kde se okno i krok mění. Údaje z tabulky je možné snadno zkopírováním vložit do libovolného statistického programu a dále s nimi pracovat.

2.2 Synchronní část

Při porovnávání variant v synchronním pohledu jsme jako výchozí distinkci zvolili rozdíl mezi psaným a mluveným jazykem, která je obecně považována

---

[4] Ačkoli v tomto ohledu nebyla provedena žádná exaktní měření, soudíme, že subjektivně pociťovaný rozdíl ve spolehlivosti údajů na frekvenční hladině 2 a 10 je větší než rozdíl mezi spolehlivostí výsledků založených na frekvencích 102 a 110.

za primární (např. Čermák 1993, Kořenský 2005: 274). Údaje o psaném jazyce jsou v celé synchronní části odvozeny od výsledku hledání v korpusu SYN2010, zatímco informace o mluveném jazyce se odvozují z korpusů Oral2006 a Oral 2008.

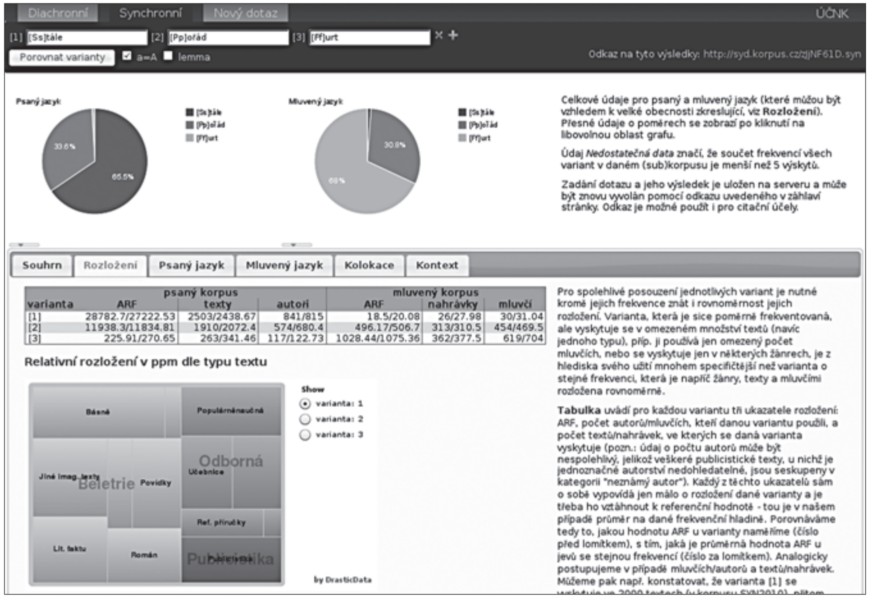
Celkové údaje pro psaný a mluvený jazyk, s kterými je uživatel konfrontován nejdříve a které můžou být vzhledem k velké obecnosti zkreslující, jsou doplněny poznámkou o tom, že zejména u méně frekventovaných jevů a jevů nerovnoměrně rozložených napříč texty je pro správnou interpretaci třeba konzultovat situaci v konkrétních subkorpusech psaného i mluveného jazyka.

K tomu slouží především oddíl Rozložení, který podává přesnější informace o jednotlivých variantách a rovnoměrnosti rozložení jejich výskytů v korpusech. Varianta, která je sice poměrně frekventovaná, ale vyskytuje se v omezeném množství textů (navíc jednoho typu), případně ji používá jen omezený počet mluvčích, nebo se vyskytuje jen v některých žánrech, je z hlediska svého užití mnohem specifičtější než varianta o stejné celkové frekvenci, která je napříč žánry, texty a mluvčími rozložena rovnoměrně.

Tabulka uvádí pro každou variantu tři ukazatele rozložení: průměrnou redukovanou četnost – ARF (Savický-Hlaváčová 2002), počet autorů/mluvčích[5], kteří danou variantu použili, a počet textů/nahrávek, ve kterých se daná varianta vyskytuje. Každý z těchto ukazatelů sám o sobě vypovídá jen málo o rozložení dané varianty a je třeba ho vztáhnout k referenční hodnotě – tou je v našem případě průměr na dané frekvenční hladině. Porovnáváme tedy například to, jakou hodnotu ARF u varianty naměříme, s tím, jaká je průměrná hodnota ARF u jevů se stejnou frekvencí. Analogicky postupujeme v případě mluvčích/autorů a textů/nahrávek. Můžeme pak např. konstatovat, že varianta 1 se vyskytuje ve 2000 textech (v korpusu SYN2010), přitom jevy na stejné frekvenční hladině (s frekvencí odlišnou maximálně o 1 %) se objevují průměrně v 1500 textech, nebo že varianta 2 je užívána jen 15 mluvčími (v korpusech Oral2006 a Oral2008), přičemž jevy podobné frekvence v průměru užívá 80 mluvčích.

Barevně jsou v tabulce pro snadnější orientaci odlišeny případy nadprůměrně rovnoměrného rozložení (zelená barva značící velkou spolehlivost naměřených údajů) a naopak případy nadprůměrně nerovnoměrného rozložení (různé odstíny pro nerovnoměrné, výrazně nerovnoměrné a extrémně nerovnoměrné rozložení), což signalizuje, že zkoumaná problematika si vyžaduje podrobnější analýzu (zkoumání variace v jednotlivých typech textů, žánrech, sociálních skupinách mluvčích apod.). Údaje v šedé barvě vycházejí z tak nízkých frekvencí výskytů, že nemají prakticky žádnou vypovídací hodnotu.

Sekce Psaný jazyk a Mluvený jazyk doplňují obecné informace prezentované úvodními grafy o detailnější rozbor distribuce variant v specifických oblastech. Uživatel tak v sekci psaný jazyk může zjistit poměr variant v rámci jednotlivých textových typů[6] a žánrů, případně i v závislosti na médiu a zdrojovém jazyku. Situace v mluveném jazyku je podrobněji specifikována na základě sociolingvistických kategorií obsažených v korpusech Oral2006 a Oral2008: pohlaví, věk, vzdělání a regionální příslušnost[7].



Odkaz na tyto výsledky: http://syd.korpus.cz/zjNF61D.syn

Celkové údaje pro psaný a mluvený jazyk (které můžou být vzhledem k velké obecnosti zkreslující, viz **Rozložení**). Přesné údaje o poměrech se zobrazí po kliknutí na libovolnou oblast grafu.

Údaj *Nedostatečná data* značí, že součet frekvencí všech variant v daném (sub)korpusu je menší než 5 výskytů.

Zadání dotazu a jeho výsledek je uložen na serveru a může být znovu vyvolán pomocí odkazu uvedeného v záhlaví stránky. Odkaz je možné použít i pro citační účely.

	psaný korpus			mluvený korpus		
varianta	ARF	texty	autoři	ARF	nahrávky	mluvčí
[1]	28782.7/27222.53	2503/2438.67	841/815	18.5/20.08	26/27.98	30/31.04
[2]	11938.3/11834.81	1910/2072.4	574/680.4	496.17/506.7	313/310.5	454/469.5
[3]	225.91/270.65	263/341.46	117/122.73	1028.44/1075.36	362/377.5	619/704

Relativní rozložení v ppm dle typu textu

Pro spolehlivé posouzení jednotlivých variant je nutné kromě jejich frekvence znát i rovnoměrnost jejich rozložení. Varianta, která je sice poměrně frekventovaná, ale vyskytuje se v omezeném množství textů (navíc jednoho typu), příp. ji používá jen omezený počet mluvčích, nebo se vyskytuje jen v některých žánrech, je z hlediska svého užití mnohem specifičtější než varianta o stejné frekvenci, která je napříč žánry, texty a mluvčími rozložena rovnoměrně.

Tabulka uvádí pro každou variantu tři ukazatele rozložení: ARF, počet autorů/mluvčích, kteří danou variantu použili, a počet textů/nahrávek, ve kterých se daná varianta vyskytuje (pozn.: údaj o počtu autorů může být nespolehlivý, jelikož veškeré publicistické texty, u nichž je jednoznačné autorství nedohledatelné, jsou seskupeny v kategorii "neznámý autor"). Každý z těchto ukazatelů sám o sobě vypovídá jen málo o rozložení dané varianty a je třeba ho vztáhnout k referenční hodnotě - tou je v našem případě průměr na dané frekvenční hladině. Porovnáváme tedy to, jakou hodnotu ARF u varianty naměříme (číslo před lomítkem), s tím, jaká je průměrná hodnota ARF u jevů se stejnou frekvencí (číslo za lomítkem). Analogicky postupujeme v případě mluvčích/autorů a textů/nahrávek. Můžeme pak např. konstatovat, že varianta [1] se

Obr. 4: Synchronní část SyD – poměr variant *stále* × *pořád* × *furt*. V horní části je základní přehled o zastoupení jevů v rámci psaného a mluveného jazyka, spodní polovina ukazuje rozložení první varianty (*stále*) v typech textů. Grafy rozložení ukazují relativní zastoupení každé z variant v jednotlivých částech korpusu. Velikost oblasti v grafu značí relativní frekventovanost varianty v dané skupině textů. Údaje o frekvenci jsou relativizovány vzhledem k celkové velikosti dané skupiny textů, což umožňuje srovnávat výsledky napříč jednotlivými skupinami textů navzdory tomu, že nejsou stejně velké.

---

[5] V případě autorů může být absolutní údaj zkreslující, jelikož velká část publicistických textů v korpusu psaného jazyka nemá jednoznačného autora (Y = autor nezjistitelný, příp. kolektiv).

[6] SyD poskytuje uživateli informaci o rozložení v textových typech ve dvou úrovních obecnosti. První úroveň zahrnuje pouze informaci o makroskupinách (v datové struktuře korpusů řady SYN se jedná o atribut txtype_group) beletrie, publicistika a odborná literatura, které jsou doplněny o korespondenci (data pocházejí z korpusu KSK-Dopisy). Druhá úroveň je podrobnější a ukazuje poměry variant v rámci jednotlivých (klasických) textových typů: NOV, COL, FAC, IMA, VER, SON, SCR, SCI, POP, TXB, ENC, ADM a PUB (viz www.korpus.cz).

[7] Sběr dat pro mluvené korpusy Oral2006 a Oral2008 neprobíhal od začátku na celém území ČR, ale pouze v jeho západní části. Údaje prezentované v tomto oddíle proto neodrážejí stav v celé ČR, ale pouze v 5 oblastech (s nestejným počtem nahrávek a mluvčích): středočeská, severovýchodočeská, jihozápadočeská, české pohraničí, česko-moravská (přechodná).

U všech grafů v synchronní části platí, že po kliknutí na libovolnou oblast grafu se zobrazí tabulka s číselnými údaji pro každou z variant: absolutní frekvence (tj. počet výskytů dané varianty v konkrétním subkorpusu), relativní frekvence v ppm (počet výskytů na milion slov), a konečně relativní frekvence v procentech, vyjadřující srovnání s ostatními variantami (součet této hodnoty u všech variant se rovná 100%). Obecná zásada rovněž uplatňovaná na všechny grafy všech (pod) oblastí současného jazyka stanoví, že výsledky jsou zobrazeny jako smysluplně interpretovatelné, jestliže součet všech variant přesáhl v daném subkorpusu kritickou mez 5 výskytů (v opačném případě se místo grafu zobrazuje upozornění, že pro daný problém nejsou k dispozici dostatečná data).

Posledním oddílem synchronní analýzy variant je rozbor jejich kolokačních profilů. Varianty, které jinak vykazují velmi podobné formální, významové i frekvenční charakteristiky, se právě v oblasti kolokability často liší.

Diagramy ukazují ke každé variantě výběr z nejdůležitějších kolokací v psaném jazyce (korpus SYN2010). Při přejezdu ukazatele myši se kolokující slova (napříč variantami) zvýrazní, což umožňuje jednoduše identifikovat kolokáty společné oběma (resp. všem) variantám. Při kliknutí na slovo se zobrazí náhodný vzorek konkordančních řádků (maximálně 25) dané varianty a kolokujícího lemmatu.

Zobrazení kolokací, tzv. *term cloud*, vyjadřuje několik hodnot popisujících pevnost a frekvenci kolokací současně. Velikost fontu je odvozena od hodnoty kolokační míry známé jako *MI/t-score* (Evert 2004: 90). Ta je definována jako menší hodnota z dvojice známých měr *MI-score* a *t-score*. Kombinuje tak výhody obou měr, kdy *MI-score* nadhodnocuje kolokace s celkově nízkou frekvencí, zatímco *t-score* neúměrně vysoce hodnotí kolokace s vysokou frekvencí.

Barva fontu (od světle modré, přes tmavě modrou až po jasně červenou) je odvozena od kolokační míry známé jako *dice* (viz Smadja 1993). Vzhledem k její konstrukci (viz souhrnné informace o různých asociačních mírách na např. www.collocations.de) – nabývá hodnot mezi 0 a 1 a není tolik závislá na frekvenci kolokace – je zajímavým doplňkem k *MI/t-score*.

Čísla v závorce za každým slovem představují absolutní frekvenci souvýskytu daného slova s hledanou variantou (s maximální vzdáleností dvě textové pozice).

Do seznamu je vybíráno až 20 kolokací s nejvyšší hodnotou *MI/t-score*, dále pak kolokace, které se objevují ve dvacítce nejdůležitějších u jiných variant. Minimální frekvence kolokátu přitom musí být alespoň 3 výskyty.

[1] [Ss]tále

často(1250) co(215) dokola(318) držet(529) hodně(2892) živý(218)
já(890) jen(446) ještě(5749) mít(2341) muset(500) my(534) něco(288)
nedohledno(33) objevovat(244) omílat(12) opakovat(311)
opakující(73) rostoucí(222) rozšiřující(39) sílící(41)
střeh(49) stejně(234) stejný(795) stoupat(226) stávat(239) tam(252)
ten(1767) více(1017) zůstávat(702) zvyšující(69)

[2] [Pp]ořád

chodit(222) chápat(116) co(340) dělat(260) dokola(674) dokolečka(28)
držet(213) hodně(209) já(1632) jak(183) jen(539) jenom(223) ještě(4408) říkat(386) mlít(73)
motat(25) mít(1822) muset(580) my(391) něco(509) někde(119) omílat(35)
opakovat(245) otravovat(37) rovně(35) samý(117) skoro(163) střeh(27)
stejně(372) stejný(455) šít(25) tam(438) ten(3051) ty(356) zůstávat(135)

[3] [Ff]urt

chodit(11) chápat(3) co(25) dělat(15) dokola(15) furt(10) já(72) jak(17) jen(16)
jenom(10) ještě(30) říkat(18) mlít(5) motat(4) mít(44) muset(17) my(18)
něco(19) někde(5) otravovat(3) samý(4) sem(12) skoro(7) stejný(13) tam(20) ten(110)
ty(15)

Obr 5: Kolokační analýza lexikálních variant *stále* vs. *pořád* vs. *furt*.

3. Výhledy do budoucna

Vedle srovnání kolokací by se dalším doplňkem komplexního pohledu na varianty měla stát kontextová analýza (Cvrček 2010). Jejím základním principem je fakt, že kontext odráží všechny podstatné rysy jednotky (formální i sémantické), a tudíž jeho zkoumáním můžeme dojít k zajímavým poznatkům o sledovaném jevu.

Vedle toho je možné na základě kontextů porovnávat jednotky a zjišťovat jejich kontextovou blízkost – předpokládáme, že vstupují-li do podobných kontextů, mají podobné funkce, formu i sémantiku (viz Cvrček 2010). V našem případě by bylo zejména zajímavé zabývat se jednotlivými zadanými variantami z tohoto pohledu a zjišťovat jejich vzájemnou blízkost v porovnání s ostatními slovy v korpusu.

Vhodným měřítkem kontextové blízkosti může být např. korelační koeficient mezi frekvencemi jednotlivých kontextů. Tyto kontexty přitom můžou být tvořeny různými jednotkami (na různých úrovních zobecnění). Nejobecnější kontext tak představují frekvence slovních druhů vyskytujících se v okolí hledané jednotky (bez ohledu na to, jestli jednotce předcházejí nebo za ní následují). Na druhé straně nejspecifičtějším kontextem by byly kontexty založené na dvojicích slovních tvarů, z nichž první předchází KWIC a druhý ho následuje. Typ kontextu a míra jeho specifičnosti přitom rozhodujícím způsobem ovlivňuje výsledky kontextové analýzy – čím specifičtější kontext, tím lepší precision celé analýzy, ale horší recall a opačně. Zároveň se ukazuje, že kontext tvořený slovními druhy zdůrazňuje shody na této úrovni (jako kontextově nejpodobnější preferuje jevy se stejným slovním druhem), kontext tvořený lemmaty se naopak zaměřuje spíše na lexikologické podobnosti atp.

Porovnávání každé z hledaných variant s celým korpusem je úkol, který dosud svojí časovou a výpočetní náročností přesahuje limity webové aplikace. Věříme však, že se záhy tento problém podaří vyřešit a budeme schopni v relativně krátkém čase uživateli nabídnout poměrně přesnou přesnou informaci o kontextové blízkosti obou variant (což může být informace zásadní pro zhodnocení jejich vzájemné zaměnitelnosti), ale také přehled kontextově nejpodobnějších slov z korpusu psané češtiny, jejichž vyhodnocení bude probíhat na nejrůznějších úrovních obecnosti kontextu.

Mezi technická desideráta celé aplikace patří především export dat do různých formátů. V případě grafů je pro publikační účely vhodné poskytnout uživateli výsledek v kvalitě umožňující tisk (optimálně v jednom z formátů pro vektorovou grafiku), v případě dat pak export do standardních (a jiných běžně používaných) formátů jako jsou CSV (comma separated values), OpenDocument formát či MS Excel, které umožní další práci s daty ve statistických programech.

4. Závěr

V kontrastu k tradičním přístupům k češtině, které se snažily jazykovou variabilitu spíše potlačovat, je tento korpusový průzkum variant zaměřen jednoznačně na jejich poznávání a funkční odlišování. Věříme, že nástroj, který se chystáme uvést do provozu, umožní uživatelům jazyka rozhodovat se při vytváření vlastních textů na základě relevantních údajů o skutečném úzu jednotlivých jednotek. Úzus odráží vždy vlastnosti dané jednotky plastičtěji než sebelepší kodifikační příručka. Věříme proto také, že tento fakt ocení i uživatelé jazyka a tato aplikace tak napomůže svým malým dílem k posunu ve vnímání jazyka směrem od předmětu regulace spíše k objektu našeho poznání.

Literatura

Cvrček, V., 2006, *Teorie jazykové kultury po roce 1945*. Karolinum. Praha.

Cvrček, V., 2008, *Regulace jazyka a Koncept minimální intervence*. NLN. Praha.

Cvrček, V. et al, 2010, *Mluvnice současné češtiny*. Karolinum. Praha.

Cvrček, V., 2010, A Contextual Approach to Parts of Speech. In *InterCorp: Exploring a Multilingual Corpus*. NLN. Praha, (s. 190-204).

Čermák, F., 1993, Spoken Czech. Varieties of Czech. Studies In *Czech Sociolinguistics*. Ed. Eva Eckert. Amsterdam – Atlanta, (s. 27–41).

Český národní korpus - DIAKORP. Ústav Českého národního korpusu FF UK, Praha. Cit. 23.05.2011, dostupný z WWW: <http://www.korpus.cz>.

Český národní korpus - SYN. Ústav Českého národního korpusu FF UK, Praha. Cit. 23.05.2011, dostupný z WWW: <http://www.korpus.cz>.

Český národní korpus - KSK-DOPISY. Ústav Českého národního korpusu FF UK, Praha 2005. Dostupný z WWW: <http://www.korpus.cz>.

Český národní korpus - SYN2005. Ústav Českého národního korpusu FF UK, Praha 2005. Dostupný z WWW: <http://www.korpus.cz>.

Český národní korpus - ORAL2006. Ústav Českého národního korpusu FF UK, Praha 2006. Dostupný z WWW: <http://www.korpus.cz>.

Český národní korpus - ORAL2008. Ústav Českého národního korpusu FF UK, Praha 2008. Dostupný z WWW: <http://www.korpus.cz>.

Evert, S., 2004, *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD dissertation, IMS, University of Stuttgart. Published in 2005

Kořenský, J., 2005, *K článku Od školské spisovnosti ke standardní češtině: reakce na výzvu k diskusi*. SaS 66, (s. 270–277).

Palková, Z., 1993, Spisovný standard jazyka v mluvené komunikaci. In *Spisovná čeština a jazyková kultura 1993*. Eds J. Jančáková, M. Komárek, O. Uličný, Filozofická fakulta UK. Praha 1995, (s. 76–80).

Savický, P. - Hlaváčová, J., 2002, Measures of Word Commonness. In *Journal of Quantitative Linguistics*. Swets & Zeitlinger, Vol. 9, No. 3, (s. 215–231).

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