ČASOPIS PRO MODERNÍ FILOLOGII 96, 2014, Č. 1, S. 9–26

Kvantitativní určení lexikálního jádra jazyka[1]

Václav Cvrček (Praha)

QUANTITATIVE DELIMITATION OF THE CORE OF A LANGUAGE

The exploitation of hapax legomena, i.e. word or lemma types which occur in a corpus only once, is usually overlooked in language description. These types cannot be systematically used for a vast majority of analyses as they do not provide a basis for any type of generalization. On the other hand, the overall number of hapaxes can be used as an indicator of the lexical periphery of the language system. This paper suggests that the ratio between the number of hapaxes and the number of all types in relation to the growing corpus size (hapax-type ratio, HTR) can be used for delimitation of the lexical core of a language. It has been shown by previous research (Fengxiang 2010) that HTR in English has the shape of a pipe or chibouque, which means that the rates of the emergence of new hapaxes and new types in the process of building a corpus differ before and after reaching a certain size. In a hypothetical small corpus (a few sentences) the hapax-type ratio will be equal to one (each word-type is also a hapax). As texts are added to the corpus (up to a few million words), the hapax-type ratio decreases (the number of new words including hapaxes is continuously increasing but the majority of added tokens are new instances of words already present in the corpus) from its maximal value (=1) to a local minimum. After reaching this turning point, extending the corpus increases the ratio because the number of hapaxes grows at a faster pace than the number of non-hapaxes (i.e. types with a frequency higher than one). This empirical finding tested on corpora of Czech and English brings us closer to the exact determination of the range of the core lexicon. Subsequently, we can deduce the approximate size of a corpus sufficient for compiling a dictionary that covers the core lexicon.

KEYWORDS

corpus, quantitative linguistics, hapax legomenon, lexicon, token-type ratio

KLÍČOVÁ SLOVA

korpus, kvantitativní lingvistika, hapax legomenon, lexikon, token-type poměr

1. ÚVOD

Vymezení toho, co můžeme považovat za jádro jazyka, je zásadní metodologická otázka nejen pro lingvistiku, ale také pro valnou část jejích aplikací jako je např. výuka nebo počítačové zpracování jazyka (NLP). Jedním z prvních výrazů snahy vyčlenit minimální objem jazykových elementů schopných pokrýt maximum komunikace je koncept tzv. Basic English (Ogden, 1930; Crystal, 1997, s. 358). Právě v tomto průkopnickém pokusu se ukázalo, že frekvence jevů (v tomto případě slov) musí hrát ve všech podobných snahách zásadní úlohu, což ostatně demonstrovali později svými

1 Tato studie vznikla v rámci Programu rozvoje vědních oblastí na Univerzitě Karlově č. P11 Český národní korpus, podprogram Český národní korpus.

pracemi i další badatelé (textové pokrytí jevů různé frekvence v angličtině sumarizují např. Francis — Kučera, 1982, nebo Waring — Nation, 1997).

Lingvistická motivace pro upřesnění velikosti jádra jazyka vychází v podstatě ze snahy zjistit, jaký rozsah popisu je přiměřený pro jazyk, který je (na rozdíl od praktických možností jakékoli deskripce) alespoň ve své potencialitě nekonečný. Vědomě nebo podvědomě vycházíme z toho, že existuje jakási konečná množina prostředků, jejímž popisem zachytíme valnou většinu relevantních rysů daného subsystému jazyka a jejíž rozsah musí být kognitivně přiměřený, aby vůbec mohl být většinou mluvčích internalizován a využíván.

Korpusový výzkum jazyka je založen na předpokladu, že rozsáhlé a reprezentativní vzorky jazyka věrně reflektují jazykovou realitu kolem nás (resp. že ji zkreslují ze všech dostupných metod nejméně). V případě lexikografie, o kterou půjde v našem příspěvku především, se domníváme, že nespecializovaný korpus je vedle informací o periferních jednotkách, které jsou kusé a nesoustavné, dostatečným zdrojem dat o jádře slovní zásoby.

Tradičním zdrojem informací o podobě slovní zásoby jsou slovníky. Srovnáme-li je ovšem s korpusy, zjistíme, že relativně často ve slovnících důležitá hesla nenacházíme (rozhodně častěji, než zjišťujeme v korpusech absenci slov ve slovníku uvedených). Tyto rozdíly nepramení pouze z náhodných opomenutí na straně slovníkářů nebo z málo precizního designu korpusů, což může vést k jejich nereprezentativnosti; na vině je zde v mnoha případech rozdílný přístup k tomu, co za jádrový prostředek považovat a co ne, tj. kde vést dělicí čáru mezi centrem a periférií v oblasti lexikonu.

Je přitom zjevné, že o jádru lze uvažovat v souvislosti s různými kritérii. Jádro je možné vymezovat s ohledem na systémovost, resp. anomálnost, tedy pomocí pojmů pražské školy jako centrum a periférie (srov. Daneš, 1965; Čermák, 2011, s. 115), nebo prizmatem úzu, tedy co je běžné a co je neobvyklé. Z dalších vymezení je možné zmínit dělení sociolingvistické a dialektologické, která vydělují části jazyka společné všem skupinám mluvčích (ať už je vymezujeme podle věku, pohlaví, vzdělání, profese nebo jinak) a všem nářečním oblastem v protikladu k prostředkům sociálně či regionálně specifickým. Zde se zaměříme z plejády nastíněných možností na jádro vymezené frekvenčně, a to na úrovni lexikonu (přesněji na rovině jednoslovných lexémů). Frekvence jevu se přitom, jak víme, se systémovostí nemusí krýt (a často nekryje, srov. nesystémové vysoce frekventované jevy jako flexi slov *být, člověk, rok*; stupňovaní *velký, dobře* apod.). Didaktický pohled na celou problematiku (Waring — Nation, 1997; Laufer, 2010) zdůrazňuje perspektivu uživatele (zejména nerodilého mluvčího). Pohled ryze technický představují NLP aplikace, které definují jádro jazyka jako množinu prostředků, které je možné použít bez ohledu na jejich textové okolí (Zhang — Huang — Yu, 2004, s. 1121).

Jelikož cílem tohoto textu je vymezit jádro pro účely lingvistického popisu, je třeba zformulovat definici vlastní. V ní lexikální jádro jazyka nepředstavuje nejmenší možnou skupinu prostředků, které jsou schopné naplňovat základní komunikační cíle, ale tu část jazyka, která by měla být popsána menším nebo středním slovníkem jazyka tak, abychom ho mohli považovat za deskriptivně přiměřený. Řádově se tedy pohybujeme mezi tisíci a desetitisíci jednotkami. Cílem výzkumu je proto najít lingvisticky opodstatněnou hranici mezi jevy jádrovými a periferními, která by byla

využitelná při popisu jazyka a umožňovala omezit rozsah lingvistických popisů při zachování jeho relativní úplnosti.

Takováto definice přísnějšího čtenáře nemůže uspokojit, protože je jen velmi obtížně operacionalizovatelná. Cílem tohoto článku nicméně není podat teoretickou definici jádra jazyka, ale poukázat na empiricky doložitelný fakt systémového předělu mezi jednotkami různé frekvence, kterému lze přidělit interpretaci delimitátora jádra jazyka a který by byl použitelný při konstruování budoucích slovníků i korpusů.[2]

Je třeba si také uvědomit, že lexikální jádro jazyka odkazuje jak k formální stránce jednotek, tak k jejich významu. V tomto příspěvku se zaměříme na formu, protože zjišťovat jádro sémantické je současnými nástroji velmi obtížné nebo dokonce nemožné. Měření budou prováděna jak na slovních tvarech, tak na lemmatech (základních slovníkových tvarech) a to především z důvodu lepší porovnatelnosti výsledků mezi různými jazyky (s nestejnou mírou flexe), ale také pro snazší srovnání těchto dvou druhů jednotek a jimi vymezených množin jádrových prostředků. Slovem tedy v tomto příspěvku rozumíme grafickou formální jednotku souvislého textu (shluk písmen oddělený z obou stran mezerami), která může mít podobu tvaru nebo lemmatu.

2. INTUITIVNÍ VYMEZENÍ JÁDRA JAZYKA

Intuitivně bychom za lexikální jádro jazyka mohli považovat ty prostředky, které se vyskytují ve všech textech (nebo ve většině textů) daného korpusu, příp. ty, které užívá majorita autorů nebo mluvčích. Výsledky měření založené na těchto premisách jsou však překvapující. Ve stomilionovém korpusu angličtiny BNC je 4049 textů (nebo vzorků textů).[3] Pouze jediné „slovo" však najdeme ve všech těchto textech, a to je tečka (.), která v korpusovém designu běžně získává samostatnou pozici jako kterékoli jiné grafické slovo. Přirozenou námitkou proti takovémuto postupu může být poukaz na to, že některé texty v BNC jsou příliš krátké. Zaměříme-li se na texty s délkou více než 1000 slov (tokenů), zjistíme, že v takto vymezených 3828 textech najdeme pouze 58 společných slov. Pokud omezíme výběr zkoumaných textů ještě více a prozkoumáme pouze ty, jejichž délka přesahuje hranici 10 000 slov, najdeme ve 2706 textech jenom 432 průnikových slov. Nahlíženo z jiné strany můžeme konstatovat, že přesně 951 slov se vyskytuje ve více než 50 % textů korpusu BNC.

Situace v češtině je obdobná. V rovněž stomilionovém korpusu SYN2010, který je svým designem z části podobný anglickému BNC, najdeme v různých textových typech celkem 2646 textů. V každém z nich po vyloučení interpunkce objevíme pouze 9 společných slov: *a, být, na, s, se, ten, v, z, za*. Ve všech textech s minimální velikostí 5000 tokenů najdeme pouze 22 slov a jenom 2126 slov se vyskytuje alespoň v polovině textů celého korpusu.

2 Vztah rozsahu korpusu, který má být základem pro slovník, a rozsahu tohoto slovníku je přitom zřejmý.

3 V rámci korpusu BNC se v zájmu větší pestrosti textů objevují i neúplné části textů (maximální délka textu nebo jeho části je 45 tisíc tokenů).

Je zjevné, že takto vymezené jádro jazyka neodpovídá ani vzdáleně potřebám lexikografické praxe. Je třeba počítat s tím, že jádro zahrnuje minimálně všechna slova synsémantická (což jsou desítky až stovky slov) a základní autosémantika (stovky, spíše však tisíce slov). V součtu je možné očekávat, že cílovým řádem pro určení jádra jazyka jsou tisíce, spíše však desetitisíce lexémů (což rámcově odpovídá výše zmíněnému rozsahu středního slovníku). Z toho je zřejmé, že rozsah jádra, o jehož vymezení se v tomto příspěvku pokoušíme, se výrazně odlišuje od rozsahu výběru jednotek, který je určován pedagogickými potřebami. Ogdenova Basic English ve svém původním návrhu zahrnovala 850 slov (Ogden, 1930). Můžeme tedy spekulovat o dvou různých vymezeních — jádro určené s ohledem na co nejmenší rozsah nutný k porozumění a k produkci a jádro vymezené pozorováním úzu a distribucí elementů v něm.

Druhý zmíněný pohled na lexikální jádro, který bude výchozí pro další úvahy, je založen na následujících předpokladech. Každý text obsahuje prvky, které je možné považovat z hlediska celého jazyka za jádrové, a prvky periferní, které jsou specifické z mnoha příčin: zásadní je vliv autora a jeho idiolektu, úzká vazba na žánr, komunikační situaci nebo téma textu či diachronní aspekt (archaismy a neologismy). Předpokládáme přitom, že každý text má svoje specifická slova, zatímco slova jádrová jsou více méně společná všem textům (ačkoli se — jak jsme viděli výše — ve všech textech nemusí realizovat). Jinými slovy, lexikální inventář dvou textů se bude lišit mnohem pravděpodobněji v elementech periferních než v jednotkách jádrových (tato úvaha už nemusí platit o celcích, které jsou menší než text — odstavcích, větách apod., u nich je pravděpodobná odlišnost jak v periferních, tak v jádrových jevech, podrobněji viz níže).

Pro účely vymezení jádra jazyka je v tomto příspěvku využit poměr hapaxů k typům (tzv. hapax-type ratio, dále HTR). Pro lepší orientaci v textu i pro sjednocení terminologie je dobré zopakovat jinak obecně známé definice použitých termínů. Každý výskyt slovního tvaru nebo lemmatu v korpusu se nazývá **token** (říkáme např. že v korpusu BNC je 111 milionů tokenů, tj. včetně interpunkce). Tokeny značíme podle notace zavedené např. v publikaci H. Baayena (2001) písmenem *N*. Pokud nepočítáme jednotlivé realizace, ale pouze různá slova, mluvíme o **typech** (korpus SYN2010 při 121 milionech tokenech obsahuje 1,7 mil. typů slovních tvarů a 786 tisíc typů lemmat). Jejich celkový počet při dané velikosti textu se značí *V(N)*. Typ je specifická abstrakce nad textem, jedná se o jednotku languovou, která nabývá vlastností, jako je frekvence, a je ve své podstatě dekontextualizovaná (Cvrček, 2013). Typem můžou být jednotky různé úrovně; slovní tvar *school* tak má v BNC 29 410 výskytů (tokenů), zatímco lemma *school* (zahrnující tvary *school, School, schools* ad.) se vyskytuje s frekvencí 52 471 tokenů.

Poslední pojem — **hapax** (z řec. *hapax legomenon*, tedy „jednou řečené“) — označuje takové typy, které se v textu nebo korpusu vyskytují právě jednou. Jejich celkový počet v textu o délce *N* tokenů se značí *V(1,N)* a platí, že suma všech typů je rovna počtu hapaxů a slov s frekvencí vyšší než jedna. Hapaxy většinou nejsou v centru pozornosti lingvistů, protože jejich využitelnost v běžných analýzách je minimálně sporná. Je-li k dispozici pouze jediný doklad daného typu (ať už jde o slovo nebo jev jiného druhu), je velmi obtížné na takovém základě provádět jakákoli zobecnění, zjišťovat jejich význam, funkci, typické užití nebo kontext. Jejich význam tkví zejména v jejich vztahu k celkové distribuci typů v textu nebo korpusu (viz např. Good-Tu-

ringův odhad užívaný jako aproximace produktivity, Baayen, 2001, s. 57, a Cvrček — Vondřička, 2013 apod.).

V každém textu je přitom hapaxů značné množství. Např. v českém překladu románu U. Eca *Jméno růže* od Z. Frýborta o celkové délce 195 tisíc tokenů je zhruba 29 tisíc různých slovních tvarů. Z toho přes 17 tisíc (zhruba 59,7 %) připadá na hapaxy.

3. VYMEZENÍ JÁDRA NA ZÁKLADĚ KORPUSOVÝCH DAT

Myšlenková konstrukce pokusu použitá pro kvantitativní určení lexikálního jádra simuluje postup při budování korpusu, v jehož průběhu měříme poměr počtu hapaxů ke všem typům (HTR). Pro názornost si můžeme celý přístup ukázat na textu povídky K. Čapka *Povětroň*. Představme si, že se jedná o první text budovaného korpusu a po přidání každé věty budeme zjišťovat aktuální stav počtu typů a hapaxů:

Prudký vítr ohýbá v nárazech stromy v nemocniční zahradě.

Po přidání první věty, která obsahuje 10 tokenů (tečka na konci je v rámci tokenizace korpusů považována za zvláštní pozici), obsahuje náš korpus 9 různých typů (předložka *v* se opakuje dvakrát), a tudíž 8 hapaxů; index HTR tedy začíná na hodnotě 8/9 = 0,89. Se zvětšováním korpusu o další věty v textu se hodnoty začínají měnit (viz tab. 1).[4]

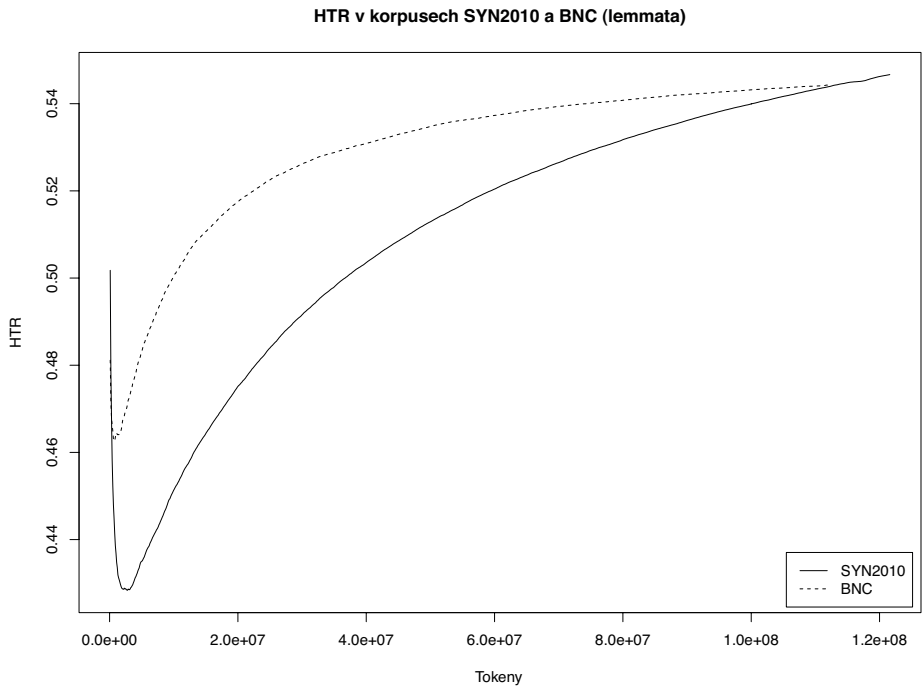
	Text/Korpus	N	V(N)	V(1,N)	HTR
1	*Prudký vítr ohýbá v nárazech stromy v nemocniční zahradě.*	10	9	8	0,89
2	*A pokaždé se ty stromy tak hrozně rozčílí, uvádí je to v zoufalství, zmítají sebou jako zástup v panice;*	32	27	24	0,89
3	*nyní se zastavují a třesou se, to nás to prohnalo, tiše, neslyšíte nic?*	49	37	32	0,86
4	*Běžme, běžme, už je to tu zas.*	59	42	35	0,83
5	*Mladý člověk v bílém plášti se klackuje zahradou a kouří si cigaretu.*	72	51	43	0,84
6	*Patrně mladý doktor;*	76	54	45	0,83
7	*vítr mu cuchá mladé vlasy a bílý plášť pleská ve větru jako prapor.*	90	64	53	0,828
8	*Jen si rvi a cuchej, divoký větře;*	99	69	57	0,826

TABULKA 1: Přírůstek typů, hapaxů a změna HTR s přibývajícím textem (zdroj: K. Čapek — Povětroň).

Vidíme, že po přidání druhé věty se HTR nemění (poměr hapaxů a typů zůstává stejný, ačkoli jejich absolutní hodnoty vzrostly). S postupným růstem korpusu ale můžeme sledovat mírný pokles HTR, který je způsoben tím, že se běžná slova začínají opakovat (každé slovo — i to sebefrekventovanější — je při prvním vstupu do korpusu hapaxem, ale pak se velmi rychle zařazuje mezi slova s frekvencí vyšší[5]). Ačkoli

4 Segmentace na věty vychází z procesu automatické tokenizace korpusů řady SYN.

5 Pro jednotky s frekvencí vyšší než 1 budeme v dalším textu používat alternativní zkrácené označení „nehapax".



GRAF 1: Poměr počtu hapaxů k počtu typů (lemmat) v korpusech SYN2010 a BNC. Hodnoty jsou výsledkem průměru 60 nezávislých náhodných sestavení textů v korpusech. Hodnoty na vodorovné ose označující velikost korpusu (N) mají formát exponenciálního čísla — např. 2.0e+07 tedy značí 2×10^7, tj. 20 mil. tokenů.

tedy počet typů i hapaxů neustále roste, hodnota HTR klesá, což značí, že počet hapaxů neroste tak rychle jako počet nehapaxů. Na druhé straně do korpusu postupně přibývají slova, která jsou poměrně vzácná a pokud se objeví, zůstávají na nejnižší možné frekvenční hladině, tj. zůstávají hapaxem. To by mohl být příklad slovního tvaru *cuchej* v poslední větě ukázky, který se vyskytuje i v 1,3 miliardovém korpusu SYN pouze jednou, a to právě v této větě z Čapkovy povídky.

Hodnoty samozřejmě výrazně ovlivňuje podoba zkoumaných jednotek. V tomto ukázkovém případě jsme se zaměřili na slovní tvary, jiné výsledky bychom získali měřením na lemmatech, které mají po mnoha stránkách k lexémům, o které při určování lexikálního jádra jde především, samozřejmě blíž. Stejně tak je třeba odlišovat hodnoty naměřené při nastavení case-insensitive (tj. bez rozlišování velikostí písmen) a case-sensitive, což je použité nastavení všech pokusů prezentovaných zde i dále.

V předkorpusové éře (nebo v dobách, kdy korpusy dosahovaly jen velmi malých rozsahů) by lingvisté pravděpodobně předpokládali, že trend poklesu HTR bude setrvalý a ustálí se okolo hladiny, která zhruba odpovídá podílu hapaxů ke všem typům v celém lexikonu. To by znamenalo, že graf vývoje HTR by měl podobu zhruba pís-

mene „L“. Výsledky měření na rozsáhlém korpusu ovšem ukazují něco jiného. Graf 1 prezentuje hodnoty naměřené na korpusech SYN2010 a BNC. Abychom eliminovali vliv pořadí dokumentů (opusů) v korpusu, vytvořili jsme šedesát nezávislých náhodných pořadí textů v obou korpusech a výsledky na nich naměřené zprůměrovali. HTR v grafu 1 byl vypočítán z množství typů a hapaxů lemmat.

Trend, který můžeme z grafu 1 vyčíst, odpovídá tomu, že na počátku sestavování hypotetického korpusu je HTR nejvyšší (poměr je velmi blízký jedné, protože dokud je korpus velmi malý, objevuje se každé slovo pouze jednou, tzn. že každý typ je zároveň hapaxem).[6] HTR následně velmi rychle klesá, až dosáhne po přidání několika milionů slov hodnoty nejnižší (lokální minimum). Toto minimum má v případě českých lemmat hodnotu 0,4284 a objevuje se v korpusu po dosažení velikosti zhruba 2,8 mil. tokenů; v průměru to odpovídá 85 026 různým lemmatům a 36 443 hapaxům. V angličtině je lokální minimum posazeno výš, což souvisí s její odlišnou typologickou charakteristikou a má hodnotu 0,4626. Nastává při velikosti korpusu zhruba 0,8 mil. tokenů, což odpovídá v průměru 30 519 různým lemmatům a 14 147 hapaxům. Od tohoto okamžiku, kdy graf dosahuje lokálního minima, se s přidáváním dalších textů HTR opět zvyšuje (i když trend je zde o poznání pomalejší než při poklesu v první fázi).

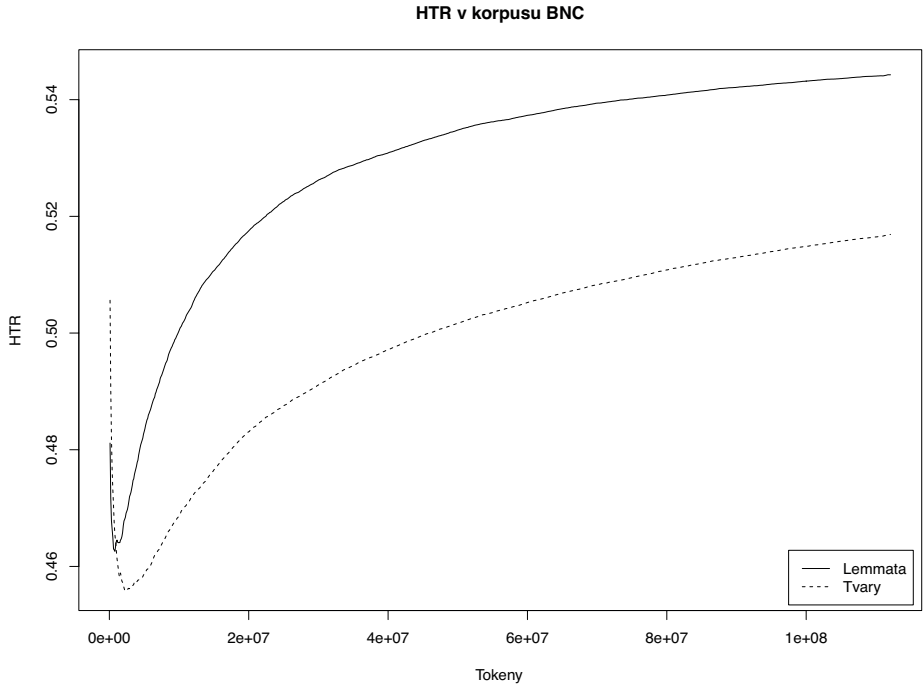
Vysvětlení klesavě-stoupavé podoby charakteristiky HTR podobné dýmce souvisí s tím, jaká slova a jakým tempem v které fázi do korpusu přibývají. Po celou dobu budování korpusu přibývají jak nová slova, tak hapaxy. Ve fázi poklesu hodnoty HTR se s větší intenzitou projevuje přírůstek nových instancí slov v korpusu už obsažených, zatímco ve fázi navazující za lokálním minimem podíl hapaxů roste rychleji než podíl nehapaxů, tedy proces přeměny hapaxů v typy s vyšší frekvencí než jedna se významně zpomaluje.

K podobným výsledkům dospěl na jiných anglických datech např. Fengxiang, 2010. I jeho grafy HTR vykazovaly specifický tvar, ke kterému budeme v následujícím textu odkazovat jako k „pipe-chart“. Odlišnost jeho přístupu k celé problematice netkví v rozdílných výsledcích, ale v interpretaci, jakou tomuto fenoménu přiřkl. Zatímco zde je průběh HTR využíván v souvislosti s určením rozsahu lexikálního jádra, přístup aplikovaný ve Fengxiang (2010) slouží zejména k optimální konstrukci korpusů a nástrojů pro účely NLP.

Podobné grafy můžeme zkonstruovat pro různé jednotky; při porovnání HTR měřeného na anglických lemmatech a slovních tvarech můžeme sledovat některé odlišnosti. První z nich je zvýšená hodnota HTR v celém průběhu křivky zobrazující hodnoty pro lemmata, která je vysvětlitelná rozdílnou mohutností inventáře lemmat a slovních tvarů, což se projevuje i na jazyce s tak chudou flexí, jako je angličtina. Druhou odlišností je umístění lokálního minima, které je nepatrně vzdálenější od počátku souřadné soustavy v případě tvarů než u lemmat (viz graf 2).

Za připomínku stojí skutečnost, že tento fenomén neintuitivního průběhu HTR nebyl pozorovatelný ještě v době vzniku prvních korpusů. Např. Brown korpus (Kučera — Francis, 1964), který obsahoval 1 milion slov, nebyl schopen o něm po-

6 Měření HTR probíhalo po úsecích o velikosti 100 tisíc tokenů. Z toho důvodu křivka HTR nezačíná na hodnotě jedna, ale níž, okolo 0.5.



GRAF 2: HTR měřený na anglických lemmatech (plná čára) a slovních tvarech (čárkovaně) v BNC (počet náhodných sestavení korpusu byl v obou případech šedesát).

dat svědectví, protože bod zvratu je pozorovatelný až při mnohonásobně větších datech (nejde jen o pozici lokálního minima, ale také o zjištění, že následná tendence je dlouhodobě rostoucí). Zároveň je třeba připomenout, že ani kvantitativní lingvistika nepředpokládala, že by existoval indikátor, který by měl v závislosti na velikosti textu či korpusu takto proměnlivé tendence (type-token poměr nebo hapax-token poměr vykazují stabilně rostoucí tendence beze změny trendu na celých datech).

Samotný bod zlomu (lokální minimum HTR) může být vysvětlen pomocí distinkce dvou typů periferních jednotek:[7] **hapaxy dočasné**, tj. ty, jejichž frekvence je jednotková v jednom textu nebo korpusu, ale zvětšením zkoumaného materiálu o další text by se velmi pravděpodobně zvýšila i jejich frekvence, a **hapaxy permanentní**, tj. ty, které jsou z hlediska jazyka skutečnou periférií. Hapaxy dočasné jsou tedy periferní pouze v určitém okruhu textů (např. lexém *sírový* se objevuje právě jednou v románu L. Fukse *Vévodkyně a kuchařka*, v odborných žánrech je ale jinak celkem běžný), naproti tomu hapaxy permanentní jsou periferní v celé zkoumané populaci (např. lexém *sítkovice* se v celém korpusu SYN2010 objevuje pouze jednou v odborném textu *Kniha o medovině* od L. Dupala). Zatímco existence dočasných ha-

7 Mohli bychom v této souvislosti uvažovat také o různých stupních perifernosti jednotek.

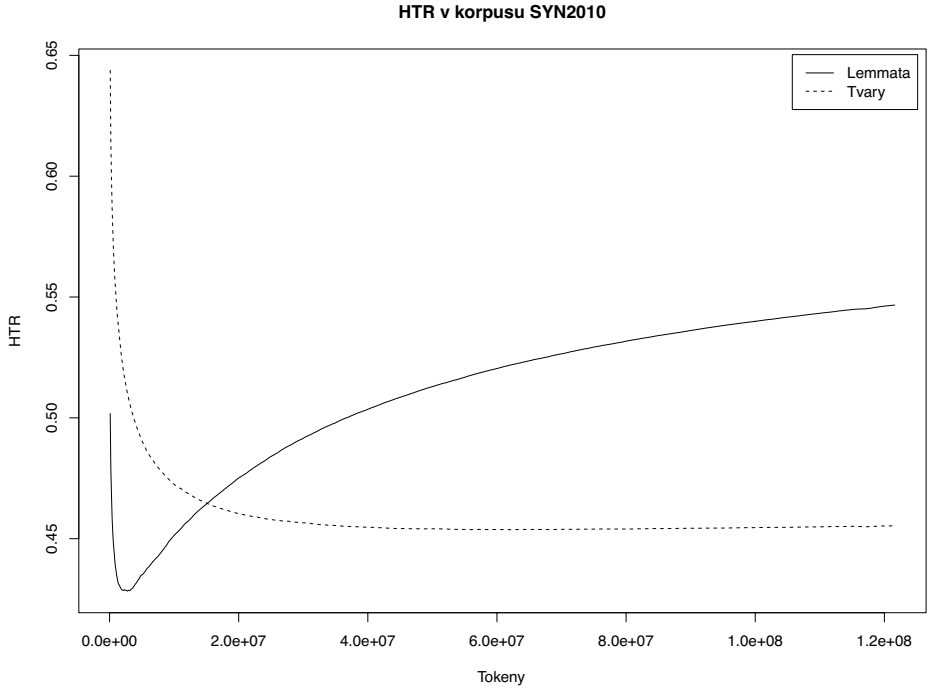
paxů je tedy důsledkem nedostatečné velikosti zkoumaného vzorku jazyka, permanentní hapaxy představují jevy řídké v celé populaci jazykových projevů. Lokální minimum v grafu HTR vzniká tedy v momentě budování korpusu, kdy drtivá většina dočasných hapaxů je v něm už obsažená a s novými přírůstky textů se pouze zvyšuje jejich frekvence a zároveň narůstá počet hapaxů permanentních. Jelikož právě o odlišení těchto dvou typů periferních jednotek při určení rozsahu jádra lexikonu jde, můžeme lokální minimum a následnou změnu trendu považovat za vhodnou aproximaci toho, jaký rozsah jádro má.

V této souvislosti je třeba zmínit, že při vymezování jádra lemmat hraje určitou roli použitý analyzátor textů (lemmatizátor). V případě běžných slov je lemmatem základní slovníkový tvar (tvarům *kůň, koně, koním* atp. je přiřazeno lemma *kůň*). V případě slov, která analyzátor nerozpozná, je lemmatem tvar samotný; to je i případ tvaru *numulitových* v korpusu SYN2010, kterému by mělo být přiřazeno lemma *numulitový*, jež ovšem není obsaženo ve slovníku analyzátoru, a tudíž je jako lemma u výskytu tohoto tvaru uvedena forma *numulitových*. Pokud by došlo k identifikaci všech tvarů adjektiva *numulitový*, zjistili bychom, že celková frekvence tohoto lemmatu by měla být 2 (vedle tvaru *numulitových* se objevuje i tvar *numulitové*). Ačkoli se zmiňované adjektivum objevuje v korpusu ve 2 tvarech, lemma *numulitový* nikdy nepřejde z kategorie hapaxů do kategorie nehapaxů, protože je nekorektně rozděleno mezi dvě lemmata.

Situace v češtině je v obecných rysech podobná angličtině, pouze s tou odlišností, že rozdíl v mohutnosti inventáře lemmat a slovních tvarů je nepoměrně větší; zatímco v angličtině je podle BNC poměr počtu různých slovních tvarů k počtu různých lemmat zhruba 1,07 (776 tisíc : 722 tisíc), v češtině je podle korpusu SYN2010 tento poměr 2,17 (1706 tisíc : 786 tisíc).[8] Z toho důvodu může křivka reprezentující slovní tvary působit jako konstantní (nerostoucí ani neklesající), nicméně její trend v druhé půlce je objektivně rostoucí, ačkoli to z grafu nemusí být na první pohled patrné. Jistý rozdíl může působit také rozdílná diversita dat v korpusech SYN2010 a BNC. Zatímco v anglickém BNC jsou zařazeny texty s maximální délkou 45 tisíc slov (v případě, že zdrojový text byl delší, byla do korpusu vtělena pouze jeho část), při budování českého korpusu SYN2010 takové omezení aplikováno nebylo, což vede k poněkud hladšímu průběhu křivky (viz graf 3).

Můžeme konstatovat, že přes nesporné typologické rozdíly mezi zkoumanými jazyky (orientační, nepublikované výsledky na korpusu současné italštiny vykazují obdobné vlastnosti), nacházíme ve všech případech shodnou tendenci k proměně průběhu funkce HTR z klesající v rostoucí. V případě obou jazyků byl vliv pořadí textů vyrušen jejich opakovaným náhodným výběrem a následným průměrováním výsledků. Zároveň se ve shodě s intuicí ukazuje, že pro různé jednotky (lemmata vs. slovní tvary) vymezují grafy HTR různě rozsáhlé množiny jádrových forem.

8 Takováto statistika je přirozeně zkreslující, protože např. lemma s celkovou frekvencí 5 výskytů nemůže mít víc než 5 různých slovních forem. Problematice objektivnějšího vyjádření míry flexe se věnuje oddíl 4 tohoto článku.



GRAF 3: HTR měřený na českých lemmatech (plná čára) a slovních tvarech (čárkovaně) v SYN2010 (počet náhodných sestavení korpusu byl v obou případech šedesát).

4. VÝSLEDKY A INTERPRETACE

Sumarizaci výsledků začněme rekapitulací užitých dat. Pro angličtinu byl využit 100 milionový korpus BNC XML edition, který obsahuje psané texty (nebo jejich části) různých žánrů (cca 90 % korpusu) a přepisy mluvených projevů (cca 10 %). Korpus SYN2010 je referenční a vyvážený korpus současné psané češtiny s následujícími poměry mezi hlavními textovými typy: 40 % beletrie, 27 % odborná literatura, 33 % publicistika. Vzhledem k celkovému počtu textů a absenci vzorkování tak můžeme SYN2010 považovat za homogennější než BNC, i když zjevně nedosahuje takové pestrosti.

Výsledky shrnující podstatné aspekty měření na obou jazycích nabízí tabulka 2.

Při interpretaci těchto výsledků je třeba mít na paměti, že ani v jednom případě nebyla zohledněna tendence (obzvláště silná v angličtině) tvořit víceslovné jednotky. Počet typů v lokálním minimu HTR tak nelze přímočaře spojovat s počtem heslových slov v případném slovníku. Vypovídací hodnota se omezuje pouze na jednoslovné formy.

Zároveň je třeba připomenout, že dva typy hapaxů — dočasné a permanentní — jsou přítomny v korpusu (v různém poměru) od počátku jeho pomyslného budování. Je tedy třeba vzít v potaz, že jádro jazyka vymezené minimem HTR zahrnuje i nemalé procento hapaxů, z nichž neznámá část jich může být permanentních. Je třeba tato

Jazyk	Jednotky	Hodnota lok. minima HTR	Velikost korpusu (N) v lok. minimu HTR	Počet typů V(N) korpusu v lok. minimu HTR	Počet hapaxů V(1,N) korpusu v lok. minimu HTR
Angličtina	slovní tvary	0,4558	2 300 000	77 255	35 247
	lemmata	0,4626	800 000	30 519	14 147
Čeština	slovní tvary	0,4537	62 000 000	1 199 248	544 125
	lemmata	0,4284	2 800 000	85 026	36 443

TABULKA 2: Výsledky měření lokálního minima HTR při náhodném přidávání textů z korpusů SYN2010 a BNC (počet náhodných sestavení, z nichž byly počítány uvedené průměrné hodnoty, byl ve všech případech 60).

čísla tedy chápat jako relativně spolehlivý odhad toho, jakých maximálních rozsahů by jádro v daném jazyce a s použitím daných jednotek mohlo dosahovat.

Při porovnávání výsledků v rámci jednoho jazyka můžeme pozorovat (už výše zmíněný) rozdíl mezi jádrem tvořeným lemmaty a slovními tvary. Vyplývá z něj, že je třeba rozsáhlejšího korpusu pro dosažení lokálního minima, měříme-li ho pomocí jednotek s bohatším inventářem (slovní tvary), než v případě pokusu na lemmatech, kterých je z podstaty věci v každém textu/korpusu míň typů. Tento rozdíl je patrný zejména v češtině, kdy je rozsah korpusu potřebný k dosažení minima HTR více než 22x větší v případě slovních tvarů než lemmat.

Z těchto pozorování bychom proto mohli odvodit specifický **koeficient flektivnosti** jazyka — jednalo by se o poměr počtu jaderných slovních tvarů k počtu jaderných lemmat. Pro češtinu vychází tento koeficient 14,1, zatímco pro angličtinu 2,53, což lze ve shodě s intuicí interpretovat následovně: české jádrové lemma v průměru zahrnuje přes 14 jádrových slovních tvarů, zatímco anglické lemma pokrývá v průměru dva a půl různých tvarů. Velmi podstatnou výhodou této konstrukce koeficientu flexe je skutečnost, že nezahrnuje jednotky periferní (ať už se jedná o archaismy, idiolekt atd., a to jak na úrovni lemmat, tak i tvarů). Umožňuje tedy odhlédnout od toho, že frekventované lexémy jsou v korpusu doloženy v mnoha svých tvarech (z nichž některé můžou být periferní), zatímco lemmata s nízkou frekvencí mají poskrovnu tvarů, ačkoli jich potenciálně můžou mít relativně velké množství. Takto konstruovaný koeficient tedy umožňuje abstrahovat od flektivních potencialit, které se ze systémového hlediska jeví jako možné, ale v praxi se realizují zřídka.

5. MĚŘENÍ VERSUS MATEMATICKÝ MODEL

Zaměřme se nyní podrobněji na fenomén tvaru křivky HTR (pipe-chart). Výše naznačenou hypotézu, že za proměnou klesajícího trendu v trend stoupavý stojí nerovnoměrná distribuce periferních jednotek v textech, můžeme ověřit porovnáním s jednoduchým matematickým modelem. Ten bude simulovat situaci rovnoměrného rozmís-

tění nejádrových prostředků, v čemž bude v jistém smyslu možná podobný představám, které o distribuci hapaxů panovaly v době před příchodem rozsáhlých korpusů.

Je zjevné, že v každém textu (a tedy i v každém korpusu) nacházíme jádrové i periferní jednotky (pro účely tohoto článku považujeme hapaxy za vhodnou aproximaci lexikální periférie). Už bylo zmíněno výše, že texty se více shodují v inventáři jádrových prostředků než v inventáři prostředků periferních. Vzájemný podíl těchto dvou druhů jednotek se různí jazyk od jazyka (závisí na mnoha faktorech: variabilitě forem v daném jazyce, inklinaci k vytváření víceslovných jednotek namísto specifických tvarů slov a odvozenin apod.). Matematický model, kterým se pokusíme aproximovat rovnoměrnou distribuci periferních jednotek, vyrůstá z intuitivního předpokladu, že v každém jazyce je poměr jádrových a periferních jednotek objektivně dán a k této hodnotě pak nutně konverguje HTR v každé kolekci textů. Mohli bychom na tomto základě předpokládat, že distribuce jádrových a periferních jednotek je v každém rozsáhlejším textu více méně podobná jejich proporci v celém jazyce. Pokud by takový předpoklad platil (a dnešní výsledky na rozsáhlých korpusech to vyvracejí), mohli bychom HTR modelovat jako vztah dvou veličin — nazvěme je M-typy a M-hapaxy, jejich vzájemný poměr pak M-HTR.

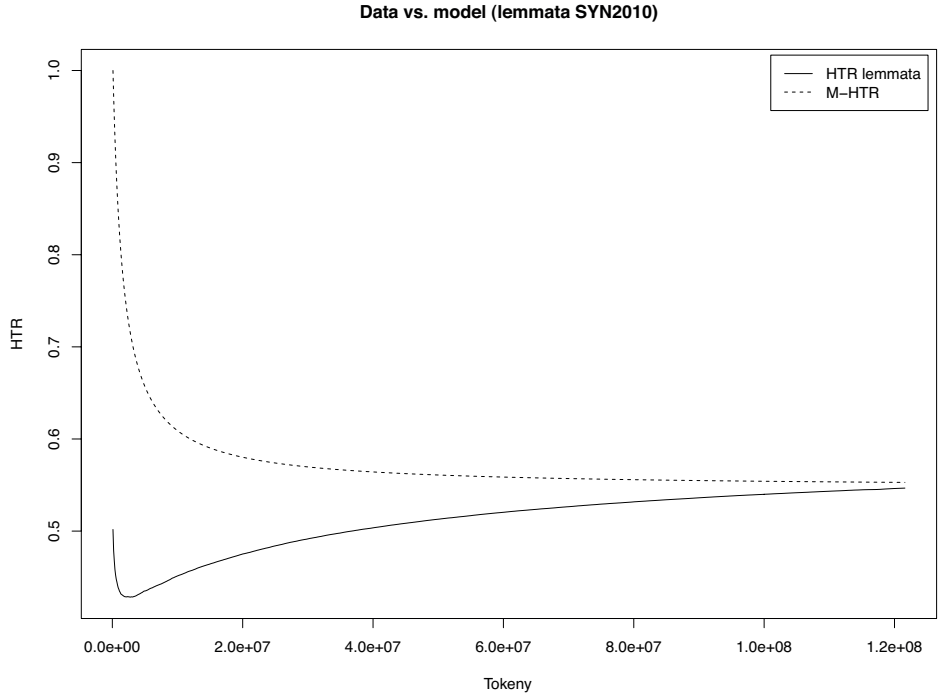
Každá z těchto veličin je definována svým přírůstkem, předpokládejme např. že M-typy přibývají tempem 100 kusů a M-hapaxy 50 kusů (obojí je vztaženo k nějaké velikosti *x* přidávaného textu). Poměr těchto dvou přírůstků je 50/100 = 0,5. Přidáme-li tedy do korpusu *x* tokenů, zvýší se počet typů o 100 a polovinu z nich budou tvořit hapaxy. Představme si nyní, že obě veličiny začínají růst z nějakého bodu, který představuje rovnovážný stav, kdy M-typy = M-hapaxy (jeho velikost není z obecného hlediska podstatná, ale pro účely tohoto textu ji stanovme např. na 1000). Na začátku je tedy poměr těchto dvou hypotetických veličin roven jedné, s každým přírůstkem se ovšem jejich vzájemný poměr mění.

Počet přírůstků (n)	**M-typy**	**M-hapaxy**	**M-HTR**
(počáteční stav)	1000	1000	1,0000
1	1100	1050	0,9545
2	1200	1100	0,9167
3	1300	1150	0,8846
4	1400	1200	0,8571
…	…	…	…
990	100000	50500	0,5050

TABULKA 3: Přírůstky M-typů, M-hapaxů a jejich vzájemný poměr (M-HTR) v matematickém modelu rovnoměrného rozmístění jaderných a periferních jednotek.

Tento model přírůstku M-hapaxů a M-typů je možné formalizovat následujícím vzorcem:

$$\text{M-HTR} = \lim_{n\to\infty} \frac{1000+50n}{1000+100n} = \lim_{n\to\infty} \frac{1000/n+50}{1000/n+100} = \frac{1}{2}$$



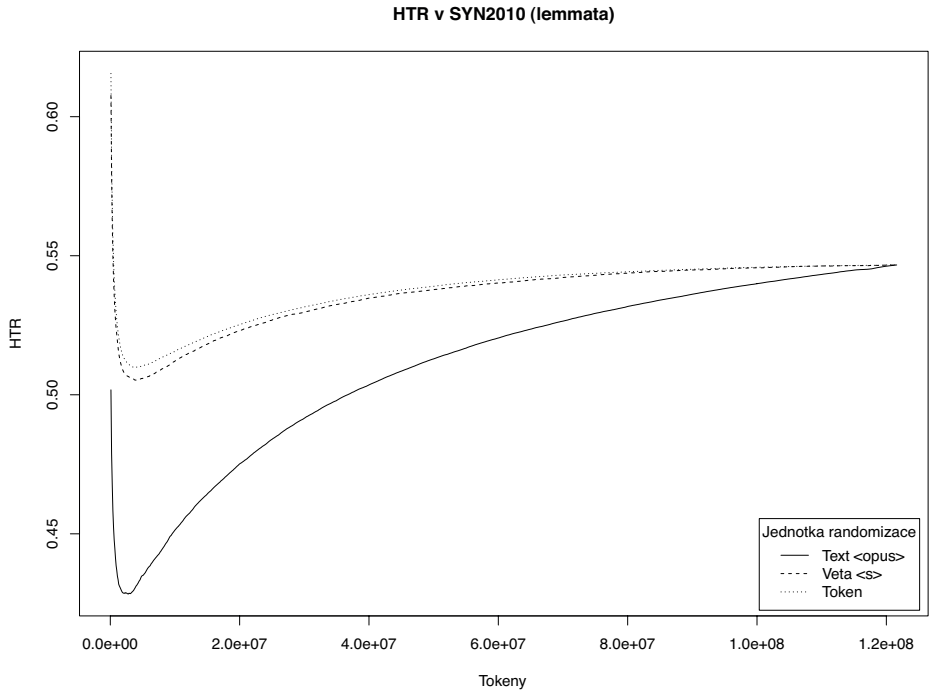
GRAF 4: Hodnoty HTR pro lemmata korpusu SYN2010 a matematický model, který simuluje stálý přírůstek M-hapaxů a M-typů (koeficienty přírůstků byly odvozeny z naměřených dat).

Zobecnění ve formě limity ukazuje, že výchozí stav s rostoucím počtem přírůstků není podstatný a naopak na významu získává podíl přírůstků. Roste-li korpus nade všechny meze, konverguje hodnota M-HTR z počátečního stavu (=1) k hodnotě, které se rovná poměru přírůstků M-hapaxů a M-typů. Porovnání matematického modelu s naměřenými daty shrnuje graf 4 (hodnoty průměrných přírůstků byly odvozeny z naměřených dat).

Jak z grafu 4 plyne, pozorovaná realita se od matematického modelu značně liší. Zatímco model je na celém definičním oboru klesající funkcí, reálný vývoj HTR má tvar dýmky (pipe-chart).[9] Odlišnost je tedy zjevně způsobena tím, že periferní jednotky nemají rovnoměrnou distribuci napříč texty.

Zatímco jádrové elementy jsou více méně ve všech textech stejné, málo frekventovaná slova (ne nutně pouze hapaxy) jsou tím, co jednotlivé texty od sebe odlišuje. Na počátku budování korpusu je každý nový typ zároveň hapaxem bez ohledu na fakt, zda se později ukáže být jádrovým nebo periferním. V procesu přidávání dalších a dalších textů se některé z typů, které v iniciálních fázích byly hapaxy, proměňují

9 Hodnoty přírůstků M-hapaxů a M-typů, které byly odvozeny z dat, nejsou podstatné; křivka matematického modelu slouží pouze k demonstraci odlišného průběhu funkce od naměřených dat.

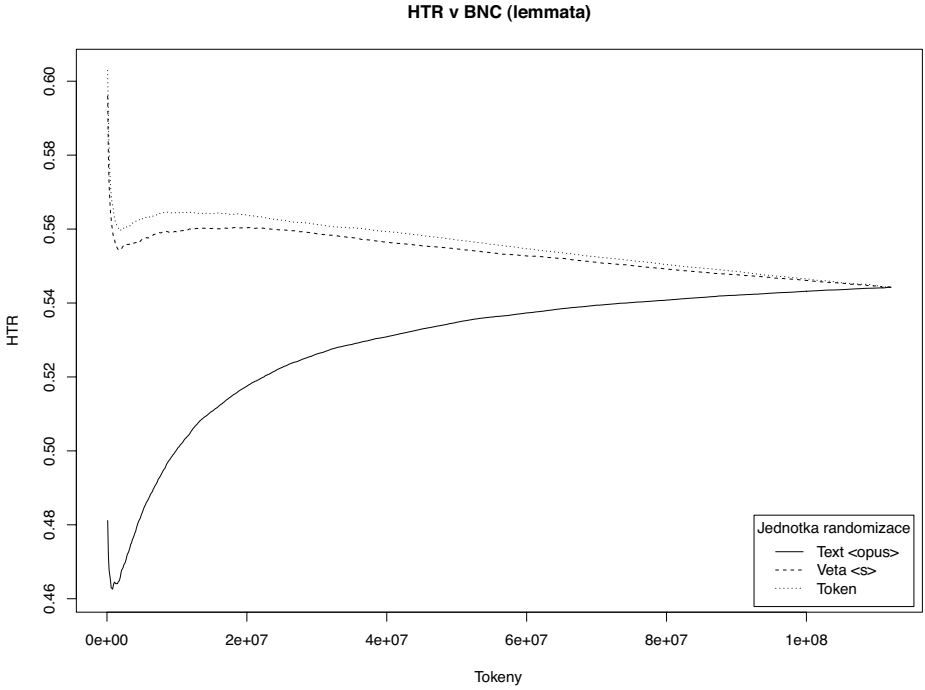


GRAF 5: Vývoj HTR měřený na lemmatech korpusu SYN2010 při randomizaci na různých úrovních — text/dokument (opus), věta a token.

v typy s vyšší frekvencí. Ty můžeme pokládat za jádrové, protože se jedná o slova, která se vyskytují v různých textech. Na druhé straně ale některé jednotky, které jsou specifické pro daný žánr, text, téma nebo autora zůstávají na nejnižší frekvenční hladině a ty je možné považovat za periferní.

Matematický model, který je popsán výše, je konstruován tak, aby distribuce periferních jednotek v něm byla rovnoměrná. Pokud pořadí jednotek v korpusu promícháme důkladněji, např. tím, že znáhodníme pořadí vět nebo tokenů (namísto celých textů, jak jsme to dělali dosud), přibližujeme se situaci v matematickém modelu. Přírůstek typů i hapaxů bude v každé fázi budování korpusu zhruba stejný a křivka vyjadřující průběh HTR se bude víc a víc podobat matematickém modelu.

Čím menší jednotku používáme pro randomizaci, tím víc graf připomíná matematický model. V případě vzorkování pomocí vět a tokenů jsou periferní jednotky, které do korpusu vstupují, promíchány ze všech oblastí jazyka (toto vzorkování totiž kombinuje všechny texty dohromady a výskyt periferních jednotek průměruje a normalizuje). Naproti tomu při vzorkování po celých textech se do korpusu nejprve dostávají periferní jednotky např. z milostného románu, později z kuchařky a dále třeba z jednoho čísla novin atp. Tím, že pořadí textů ponecháváme náhodné a výsledky zprůměrujeme, získáváme obecný trend.



GRAF 6: Vývoj HTR měřený na lemmatech korpusu BNC při randomizaci na různých úrovních — text (opus), věta a token.

Rozdíl mezi měřením HTR na datech s náhodným pořadím textů, vět nebo tokenů nespočívá ale pouze v ne/rovnoměrné distribuci hapaxů v každém přírůstku. Pramení také z nestejného rozložení jednotek s nízkou frekvencí, které nejsou hapaxy (mají frekvenci dva, tři nebo čtyři výskyty). Tato slova jsou většinou specifická pro daný žánr nebo téma, a proto se objevují pouze v určité oblasti korpusu. Je třeba si uvědomit, že téměř 30 % českých slovních tvarů, které se v celém korpusu vyskytují právě ve dvou výskytech (tzv. dis legomena), se nacházejí v rámci jednoho textu, podobně zhruba 39 % slovních tvarů s frekvencí 3 se vyskytuje v méně než třech dokumentech apod.

Pokud tedy rozšiřujeme korpus o celé texty, velká část těchto málo frekventovaných jednotek (ovšem s frekvencí vyšší než jedna) se v průběhu přidání dokumentu ihned přemění v nehapax. Vzorkujeme-li ovšem korpus po větách nebo dokonce po tokenech, nemusí k jejich přeměně dojít ani po přidání textů o celkové délce několik milionů slov.

Když tedy zcela promícháme pořadí vět nebo slov v korpusu, velmi ztížíme rozpoznání momentu, ve kterém většina jádrových typů už přestala existovat v jednom výskytu, a nové typy, které se do korpusu s dalšími texty dostávají, zůstávají mnohem častěji hapaxy, kvůli své specifičnosti a omezenosti použití.

Dalším pozorovatelným efektem změny jednotky, pomocí které korpus randomizujeme, je iniciální hodnota HTR. Zatímco v případě vzorkování po textech je výchozí

hodnota (po dosažení 100 tisíc tokenů) kvůli relativně nízké variabilitě typů zhruba okolo 0,513, v případě vzorkování po jednotlivých tokenech je to při stejné velikosti korpusu 0,619. I to je způsobeno tím, že každý text je vždy z hlediska použitého slovníku (počtu typů) homogennější než náhodný vzorek slov nebo vět o stejné délce. Situace v angličtině (viz graf 5) je analogická české s tím, že její výsledky v případě vzorkování po tokenech i po větách připomínají matematický model ještě víc než data z češtiny.

Odlišnosti mezi českými a anglickými výsledky můžeme přičíst na vrub jednak morfologické bohatosti a celkové variabilitě forem v češtině. Zatímco v angličtině je většina variability odvozena z lexikonu (rozdílné formy prezentují povětšinou rozdílné významy), v češtině za variabilitou repertoáru forem stojí morfologie. Specifické způsoby derivace, které jsou využívány zřídka, způsobují, že lemmata takto vytvořená jen obtížně překračují hranici dvou výskytů, což je tendence, která v angličtině nepůsobí zdaleka tak silně.

7. ZÁVĚRY

Neintuitivní průběh HTR na datech rozsáhlých korpusů vyjevuje mnoho podstatných informací o jazyce a korpusech; nejpodstatnější z nich je zřejmě demarkační linie (lokální minimum HTR), která by mohla být empiricky založeným vodítkem pro stanovení rozsahu jádra lexikonu. Je přitom třeba zdůraznit, že pokus popsaný výše není schopen odhalit, které prostředky za jádrové považovat a které ne, ale pouze poukázat na rozsah centrálních jevů v lexikonu. Složení tohoto jádra je třeba odhalovat jinými metodami; jeví se přitom jako velmi pravděpodobné, že frekvence, příp. jiné, sofistikovanější ukazatele četnosti jako ARF (Savický — Hlaváčová, 2002), budou v těchto oblastech hrát zásadní úlohu.

Vymezení rozsahu jádra je také prvním krokem při úvahách o velikosti korpusu, který je potřebný pro řešení různých druhů výzkumných úkolů. Pokud bychom se rozhodli sestavit korpus, který bude obsahovat všechny jádrové prostředky (jistotu mít nikdy nelze, ale s velkou pravděpodobností), musíme počítat s násobky rozsahu jádra. Pokud vymezíme jádro jako 85 tisíc nejfrekventovanějších českých lemmat, které jsou obsaženy už v korpusu o velikosti zhruba 3 mil. tokenů, musíme počítat s minimální velikostí korpusu alespoň 30krát (spíše však 100krát) větší, abychom tyto jádrové prostředky mohli s rozumnou mírou jistoty analyzovat a popsat.

Tvar HTR funkce (pipe-chart) zároveň naznačuje, že existují dosud neprozkoumané rozdíly mezi textem a jazykem. Pokud existují fenomény, jejichž trend se s překročením určité hranice délky textu rapidně mění, je třeba v tomto světle revidovat i některé závěry kvantitativní lingvistiky. Výsledky měření, které jsou provedeny na nedostatečně rozsáhlých datech, můžou být tímto fenoménem ovlivněny, a nejsou tedy schopné vypovídat o celém jazyce, ale pouze o textech, na nichž bylo měření prováděno.

Výsledky uvedených měření bychom proto s jistou mírou nadsázky mohli interpretovat jako specifický případ distinkce parole-langue. Klesající část grafu HTR by odpovídala vlastnostem textů, zatímco část rostoucí vypovídá o celém jazyce (langue)

nebo nějaké jeho části (představuje totiž vlastnosti korpusu, který je natolik rozsáhlý, že eliminuje idiosynkrazie jednotlivých autorů nebo textů). Zatímco úvodní hodnoty HTR reflektují vlastnosti textů, na nichž je měřen, hodnoty ke konci definičního oboru představují vlastnosti jazyka, zejména podíl jádrových a periferních jednotek.

Při srovnávání češtiny a angličtiny se potvrdila očekávaná typologická rozdílnost jazyků a jednotek v něm obsažených. Navržený koeficient flexe, který zohledňuje pouze jádrové prostředky (lemmata i slovní tvary) by mohl sloužit jako základ pro efektivní kvantitativní srovnávání jazyků. Je přitom třeba znovu připomenout, že průzkum zahrnující pouze jednotlivá slova (a nikoli víceslovné jednotky) je do značné míry zplošťující. Jelikož je ovšem repertoár bigramů či trigramů slov mnohonásobně větší než množina jednotlivých slov, bude třeba do zkoumání těchto jevů zapojit rozsáhlejší korpusy než SYN2010 a BNC.

Specifický průběh HTR, který je závislý na dosažení určité nemalé hodnoty rozsahu korpusu, nás zároveň upozorňuje na důležitý aspekt celé problematiky kvantitativního a korpusového zkoumání jazyka. Při formulování závěrů bychom se zvýšenou pečlivostí měli dbát na to, abychom se vyjadřovali pouze o těch aspektech a oblastech jazyka, které jsme skutečně schopni pozorovat a uchopovat. Ve světle předložených zjištění bychom neměli zapomínat na to, že budoucí korpusy, které svým rozsahem mnohonásobně předčí to, s čím můžeme pracovat dnes, můžou vypovídat o fenoménech, které si v současnosti nejsme schopni ani představit.

POUŽITÉ KORPUSY

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Dostupný z: http://www.natcorp.ox.ac.uk/

Czech National Corpus — SYN2010. Institute of the Czech National Corpus FF UK, Praha 2010. Dostupný z: <http://www.korpus.cz>.

LITERATURA

Baayen, H. R. (2001): *Word Frequency Distributions*. Dordrecht/Boston/London: Kluwer Academic Publishers.

Cvrček, V. (2013): *Kvantitativní analýza kontextu*. Praha: Nakladatelství Lidové noviny.

Cvrček, V. — Vondřička. P. (2013): Nástroj pro slovotvornou analýzu jazykového korpusu. In: *Gramatika a korpus 2012*. Hradec Králové: Gaudeamus.

Crystal, D. (1997): *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Čermák, F. (2011): *Jazyk a jazykověda*. Praha: Karolinum.

Daneš, F. (1965): Vztah „centra" a „periferie" jakožto jazykové univerzálie. *Jazykovědné aktuality*, 2, s. 1–6.

Fengxiang, F. (2010): An Asymptotic Model for the English Hapax/Vocabulary Ratio. *Computational Linguistics*, 36/4, s. 631–637.

Francis, W. N. — Kučera, H. (1964): *Brown Corpus Manual. Brown University*. Providence (dostupné z http://icame.uib.no/brown/bcm.html)

Francis, W. N. — Kučera, H. (1982): *Frequency Analysis of English Usage*. Boston: Houghton Mifflin Company.

Laufer, B. (2010): Lexical threshold revisited: Lexical text coverage, learners' vocabulary

size and reading comprehension. *Reading in a Foreign Language*, 22/1, s. 15–30.

Ogden, C. K. (1930): *Basic English. A General Introduction with Rules and Grammar*. London: Kegan Paul.

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

Savický, P. — Hlaváčová, J. (2002): Measures of Word Commonness. In: *Journal of Quantitative Linguistics*, 9, s. 215–231.

Waring, R. — Nation, P. (1997): Vocabulary size, text coverage, and word lists. In: N. Schmitt — M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press, s. 6–19.

Zhang, H. — Huang, C. — Yu, S. (2004): Distributional consistency: A general method for defining a core lexicon. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC2004), s. 1119–1222.

Václav Cvrček | Ústav Českého národního korpusu, FFUK | nám. Jana Palacha 2, 116 38 Praha 1
vaclav.cvrcek@ff.cuni.cz