UNIVERZITA PALACKÉHO V OLOMOUCI

Filozofická fakulta

Katedra obecné lingvistiky

Kvantitativní analýza čínského textu zaměřená na zkoumání čínského slova určeného na základě čínské abecedy Hanyu Pinyin Fang'an

A quantitative analysis of Chinese written text aimed at examining Chinese words determined on the basis of the Chinese alphabet Hanyu Pinyin Fang'an

DISERTAČNÍ PRÁCE

Mgr. Lenka Matoušková

Studijní program: Obecná jazykověda a teorie komunikace

Školitel: Mgr. Martina Benešová, Ph.D.

Olomouc 2022

Prohlašuji, že jsem tuto disertační práci vypracovala samostatně na základě uvedené literatury.

V Olomouci dne 28.08.2022 .....

Mgr. Lenka Matoušková

Děkuji Mgr. Martině Benešové Ph.D. za odborné vedení, cenné rady a konzultace, díky kterým mohla tato práce vzniknout.

Anotace

Texty psané čínským znakovým písmem jsou jednolité a čínské znaky stojí vedle sebe bez mezer. Hranice slov tak nejsou v čínských textech vyznačeny. Předkládaná práce se zabývá čínským ortografickým slovem, které vznikne přepisem do abecedy *Hanyu pinyin fang'an* dle normy GB/T 16159–2012 jako část textu oddělená mezerami. Výzkum ověřuje, zda takto definované slovo odpovídá ekonomizujícím pravidlům v jazyce. Původní hypotéza stanoví, že čínské ortografické slovo je v souladu s jazykovou ekonomií, a proto je takto vymezené ortografické slovo možné považovat za obecně platnou definici slova v čínštině. Ortografické slovo je ověřováno prostřednictvím shody s Menzerath-Altmannovým zákonem na třech jazykových úrovních: U3 *slovo* > *slabika* > *grafém*, U2 *klauze* > *slovo* > *slabika* a U1 *souvětí* > *klauze* > *slovo*. Bylo zvoleno pět výběrových souborů: tři povídky psané v čínských znacích a dva výběrové soubory psané originálně v pinyinu, u kterých byl ponechán původní způsob segmentace. Výsledky ukazují, že ortografické slovo vymezené dle normy neodpovídá ekonomizujícím pravidlům a hypotéza tak nebyla potvrzena. Výrazně lepší výsledky byly získány u jednoho z výběrových souborů psaného originálně v pinyinu, což naznačuje, že pro projevení se ekonomizujících jazykových pravidel by segmentace čínského slova měla zahrnovat spojování vybraných funkčních slov se slovy plnovýznamovými. Dále bylo zjištěno, že texty nejsou vždy v souladu s ekonomičností jazyka a Menzerath-Altmannův zákon reaguje velmi dynamicky na různé alternace textových produkcí, například při užívání přímých řečí, uvozovacích vět a také při nahrazování osobních jmen zájmeny. Kromě toho má na platnost Menzerath-Altmannova zákona velký vliv i frekvence pozorování. Ke konci práce je nastíněn směr navazujícího výzkumu.

Klíčová slova

Kvantitativní lingvistika, Menzerath-Altmannův zákon, psaná čínština, abeceda Hanyu pinyin fang'an, ortografické slovo, jazykové jednotky

Obsah

Ediční poznámka

V práci používáme zjednodušené čínské znakové písmo. Čínské termíny převádíme do abecedy pinyin (Hànyǔ Pīnyīn Fāng'àn 汉语拼音方案), což je oficiální norma pro přepis výslovnosti čínských znaků do latinky. Nejprve uvádíme podobu bez tónů, za kterou následuje závorka s převodem do pinyinu s tóny a dále jsou uvedeny čínské znaky. Slova a slabiky zapsané v přepisové abecedě pinyin spolu s uvedením tónových značek jsou pro přehlednost uvedeny jiným typem písma Courier New (s výjimkou samotného slova pinyin a osobních jmen).

Čínština v této práci odkazuje na moderní standardní čínštinu, tedy putunghua (pǔtōnghuà 普通话), v případě psané čínštiny máme na mysli současnou podobu psané čínštiny – xiandai baihua (xiàndài báihuà 现代白话).

Úvod

Výzkum současné čínštiny pracuje s pojmem slovo, avšak přesná definice této jazykové jednotky stále chybí. Kvůli povaze čínského textu, který je soudržným útvarem, ve kterém jednotlivé znaky přiléhají těsně k sobě, není možné detekovat slovo jako je tomu například u jazyků, které používají hláskové písmo. Mnoho čínských koncepcí přistupuje k čínskému slovu jako k jednotce, která je uvedena ve slovníku, aniž by však slovo bylo přesně definováno. Třebaže existují výzkumy čínského slova, většinou se zabývají jeho délkou v historickém vývoji nebo frekvencí slov. Přesná definice slova bývá v těchto pracích často opomíjena. Kromě lexikálního přístupu však existují další možné způsoby, jak lze čínské slovo vymezit. Jsou to například syntaktické, fonologické, morfologické případně další způsoby. V naší práci se zaměříme na zkoumání ortografického slova, které vznikne po převedení textu psaného čínskými znaky do abecedy pinyin podle normy GB/T 16159–2012 *Základní pravopisná pravidla pinyinu* a ortografické slovo je tak vytvořeno jako část textu, která je ohraničena mezerami. V této kapitole nejprve představíme krátký vhled do zkoumané oblasti, seznámíme čtenáře s cílem, otázkami a významem našeho výzkumu.

Práce vychází z již provedených výzkumů, které zkoumaly obecnou platnost Menzerath-Altmannova zákona (MALu[1]) na výběrových souborech psaných v čínských znacích. Prvotní fáze výzkumu byla zaměřena na ověření obecné platnosti MALu na textech psaných v čínských zjednodušených znacích. Výsledky předešlého výzkumu přinesly zjištění, že se zákon neprojevil na všech jazykových úrovních a pravděpodobnou příčinou se zdálo být, že v hierarchii jazykových jednotek chybí další jazyková jednotka, která je větší než znak, ale menší než parcelát (tj. část souvětí vymezena interpunkčními znaménky). Jako nejpravděpodobnější jazykovou jednotkou, která chybí v řetězci jednotek, se zdálo být *slovo* (Motalová a Spáčilová 2013; Motalová a Spáčilová 2014; Motalová et al. 2016; Matoušková a Motalová 2015). Stejné výsledky jsme získali i v rámci následujícího experimentu, který ověřoval platnost MALu na různých čínských překladech básně *Raven* od autora Edgara Allana Poea, a také u dalšího výzkumu, který ověřoval platnost MALu na textu psaného v čínských tradičních znacích (Matoušková 2016).

[1] Zkratka MAL v naší práci odkazuje na zkratku Menzerath-Altmannova zákona v anglickém jazyce, tj. Menzerath-Altmann Law z důvodu zažitého používání v jiné literatuře

Poznamenejme, že kromě uvedených experimentů se kvantitativně-lingvistickým výzkumem zabývaly i kolegyně Denisa Schusterová Vicherová a Jana Kovaľová (Ščigulinská), které zkoumaly platnost MALu na čínských mluvených textech. Denisa Schusterová Vicherová poté ověřovala společně s Terezou Motalovou platnost MALu na psaných výběrových souborech a stejně jako v případě našeho výzkumu došly ke stejným závěrům, že v hierarchii jazykových jednotek pravděpodobně chybí jazyková jednotka *slovo* (Motalová a Schusterová 2016).

V návaznosti na provedené výzkumy je předmětem této práce výzkum čínského slova. Jak jsme již zmínili výše, ke slovu je možné přistupovat různými způsoby a každý z nich může být validní. V této fázi výzkumu jsme se rozhodli zaměřit se na výzkum ortografického slova. Na první pohled se zdá, že ortografické slovo v čínštině neexistuje, protože znaky k sobě těsně přiléhají. Avšak text psaný v čínských znacích je možné převézt do latinky, tj. do abecedy pinyin, a tím je možné získat text, ve kterém jsou jednotlivá slova oddělena mezerami. Převádění textu z čínských znaků do abecedy pinyin není čistě náhodné, ale musí sledovat daná pravidla.

Čínský systém pinyin je latinkový zápis čínštiny, který se v roce 1958 stal oficiální přepisovou abecedou v ČLR. I když původním záměrem reforem čínského písma bylo čínské znaky nahradit latinizovaným systémem, postupně se od této myšlenky upustilo a pinyin měl sloužit pouze jako nástroj pro zachycení standardní výslovnosti znaků (Su 2001). Také dokument *Hanyu Pinyin Fang'an* (Hànyǔ Pīnyīn Fāng'àn 汉语拼音方案) *Schéma čínské fonetické abecedy*, který v roce 1958 kodifikoval pinyin, vyjmenovává pouze izolované slabiky bez uvedení ortografických pravidel a nepočítá tak s pinyinem jako písmem rovnocenným čínským znakům. Pinyin se po svém zavedení ovšem začal objevovat na různých místech jako jsou bankovky, orientační tabule, tituly knih, oficiální dokumenty při styku se zahraničím, učebnice určené pro cizince i pro výuku čínských dětí a jeho pozice se tak začala velmi upevňovat. Z toho důvodu postupně začaly vycházet různé normy zabývající se úpravou zápisu v pinyinu, jako je například výše zmíněná norma GB/T 16159–2012, která určuje ortografická pravidla, dle kterých mají být přepisována slova a slovní spojení.

Širší celospolečenské zavádění pinyinu, v dnešní době zejména v oblasti počítačových věd, přivedlo různé vědce na myšlenku, že pinyin je vhodnějším způsobem zápisu čínštiny a začaly se ozývat hlasy pro zavedení obou písem jako paralelních, tj. digrafie (Feng a Yin 2000; Su 2001), někteří badatelé dokonce zacházejí dál a vybízejí k nahrazení čínských znaků pinyinem (Hannas 1997). Protože pinyin je v současné době

nepostradatelný a napomáhá například k šíření gramotnosti, podporuje kulturní výměnu mezi Čínou a jinými zeměmi a urychluje rozvoj informačních technologií, je třeba jeho pozici neustále upevňovat i mezi Číňany. Důsledné členění textu při převodu znaků do pinyinu je zásadní, protože při nedodržování pravidel může být text nesrozumitelný a nejednoznačný (viz Hu 2005).

I když není pravděpodobné, že v dohledné době budou čínské znaky nahrazeny pinyinem, pinyin je nepostradatelnou součástí současného moderního čínského jazyka a jeho výzkum si jistě zaslouží pozornost. Z toho důvodu je předmětem této práce právě zkoumání ortografického slova, které vznikne po převodu textu psaného v čínských znacích do abecedy pinyin.

Výzkum se bude zabývat otázkou, zda jsou navrhovaná ortografická pravidla uvedená v normě GB/T 16159–2012 *Základní pravidla pro ortografii čínské abecedy pinyin* nejvhodnějším způsobem zaznamenávání čínských slov v pinyinu a zda takto definované slovo odpovídá ekonomizujícím pravidlům v jazyce. Klademe si otázku: Je možné tímto způsobem vymezené ortografické slovo považovat za obecně platnou definici slova v čínštině? Je zvolený způsob segmentování textu na slova vhodný z pohledu obecně platných ekonomizujících jazykových zákonů?

Naší hypotézou je, že způsob členění textu na slova dle této normy odráží ekonomizující pravidla, protože norma vznikla na základě důkladné práce čínských lingvistů a jedná se tak o promyšlený způsob členění čínského textu s jasně danými pravidly, a proto by ortografické slovo mělo být v souladu s jazykovou ekonomií a mělo by odrážet obecnou definici čínského slova.

Abychom mohli nalézt odpovědi na tyto otázky a ověřit naši hypotézu, je třeba do výzkumu zahrnout metody, které umožňují detailní průzkum reprezentativního vzorku. Tuto možnost nabízí nástroje kvantitativní lingvistiky, díky kterým získá výzkum na exaktnosti, objektivitě a opakovatelnosti. Metody kvantitativní lingvistiky nestaví na dojmech a domněnkách, ale závěry vyvozují na základě získaných empirických dat. Kvantitativně-lingvisticky zaměřený výzkum je propojen s výzkumem kvalitativním, kdy na základě získaných empirických dat filolog interpretuje výsledky. Pro ověření našeho výzkumu jsme se rozhodli použít MAL, protože tento zákon odráží obecnou jazykovou vlastnost, známou jako ekonomičnost. MAL umožňuje jazykovou jednotku *slovo* ověřit nejenom jako samostatně stojící jednotku, ale je možné ji ověřovat i ve vztahu k ostatním jazykovým entitám na nižší či vyšší jazykové hladině. Můžeme tak na ni nahlížet jako na jednotku, která je součástí celého řetězce.

Výsledky této práce by mohly být nápomocné při formulován obecné definice čínského slova.

Předkládaná práce je rozdělena do třech částí. První část práce představuje užití kvantitativních metod v Číně a dále seznamuje čtenáře s kvantitativním výzkumem čínštiny mimo Čínu. Pozornost je věnována zejména Menzerath-Altmannově zákonu. Druhá část práce přibližuje metodologii. V rámci této části jsou nejprve zvolena kritéria pro volbu výběrových souborů, kde je podrobněji charakterizován čínský psaný a mluvený jazyk, dále je přiblížen vývoj fonetických přepisů čínštiny a ostatní požadavky na výběrové soubory. V rámci metodologie jsou představeny zvolené výběrové soubory a jazykové jednotky *grafém* < *slabika* < ***slovo*** < *klauze* < *souvětí*, za nimiž následuje příklad způsobu segmentace. Hlavní část práce je členěna na tři oddíly, z nichž každý oddíl se zabývá danou jazykovou úrovní. U každé jazykové úrovně jsou uvedeny tabulky s výpočty, za nimi následují grafické vizualizace a diskuze k získaným výsledkům.

Přesto je třeba získané výsledky interpretovat opatrně a mít na paměti, že i tato práce má svá omezení. Prvním z nich může být výběrové zkreslení, protože výběrové soubory nemusí odrážet celou populaci. Další potenciální omezení této práce tkví v omezená časové kapacitě, proto v závěru práce nabízíme další směry, kterými by se měly ubírat navazující experimenty.

1 Užití kvantitativních metod v lingvistice

Naše práce navazuje na kvantitativně-lingvistický výzkum, který měl za cíl ověřit platnost MALu na textech psaných čínskými znaky, a nadále budeme ověřovat hypotézy, které vyplynuly z předešlých experimentů. Nejprve se však krátce zastavme u historie kvantitativní lingvistiky, tedy oboru, do kterého spadá náš výzkum. V této kapitole se nebudeme zabývat celosvětovým vývojem kvantitativní lingvistiky, jelikož je to velmi obsáhlé téma, ale zaměříme se konkrétněji na vývoj matematické lingvistiky v Číně a tamější osobnosti současné kvantitativní lingvistiky, protože předmětem této práce je právě čínský jazyk. Chvíli se zastavíme také u osobností, které se zabývaly kvantitativním výzkumem čínštiny, ale nejsou čínského původu.

Než se však dostaneme k historickému vývoji, krátce si přibližme, co kvantitativní lingvistikou myslíme a jakým způsobem by měl kvantitativní výzkum v lingvistice vůbec vypadat. Toto odvětví lingvistiky má své počátky v druhé polovině 19. století, k jejímu prudšímu rozvoji však došlo až v druhé polovině 20. století. Oproti tradičnímu pojetí ve filologii se kvantitativní lingvistika snaží přiblížit exaktním vědám tím, že aplikuje kvantitativní metody na jazyk a pomáhá objasnit, jak spolu souvisí jazykové jevy. Kvantitativní lingvistika, stejně jako každá jiná věda, by měla formulovat hypotézy, protože „veda, ktorej chýbajú hypotézy, je protoveda a veda, ktorej hypotézy sú netestovateľné, je pseudoveda. Vo filologických vedách existujú dodnes poddisciplíny, ktoré sa uspokoja s tým, že rozmnožujú batériu pojmov, vytvárajú množstvo „-izmov“ a „-ém“ na *opis* a *klasifikáciu* javov a žijú v domnienke, že vytvárajú teóriu“ (Wimmer 2003, s. 13). Kvantitativní výzkum pracuje s daty, které musejí být určitým způsobem kvantifikovatelné. To znamená, že je možné je převést na čísla, tabulky a grafy a dále je statisticky zpracovávat (Rasinger 2013, s. 10). Kvantitativný analýza má čtyři základní cíle (Johnson 2008, s. 3):

1. Redukce dat: ve smyslu shrnutí trendů, zachycení společných aspektů apod.
2. Odvozování: zobecnění reprezentativního vzorku
3. Zjištění vzájemných vztahů
4. Zkoumání procesů, které mohou mít základ v pravděpodobnosti

Tento obor filologie neustále zrychluje svůj vývoj a filologii přibližuje teorii systémů, synergetice, fyzice a exaktní filozofii. Využití kvantitativních a matematických metod v lingvistice pomáhá zpřesnit a prohloubit výzkum a umožňuje dedukci a výstavbu teorie (Wimmer 2003, s. 13, 15).

Na rozdíl od Johnsona Rasinger zdůrazňuje, že kvantitativní výzkum by oproti kvalitativnímu výzkumu, který je založen na indukci (tzn. teorie se odvozuje na základě získaných výsledků), měl probíhat deduktivně, tzn. na základě již známé teorie vystavujeme hypotézy, které se nadále snažíme dokázat nebo vyvrátit pomocí empirických pozorování (Rasinger 2013, s. 11). Stejně jako v jiných vědách by měl kvantitativní výzkum probíhat v jednotlivých krocích:

1) **Formulování kvalitativní hypotézy** – obvykle ji formuluje filolog (ne matematik) verbálně na základě vlastních zkušeností, znalostí a relevantnosti problému v dané oblasti (Wimmer 2003, s. 15). Základním předpokladem, že hypotéza musí být empiricky relevantní a testovatelná (Köhler a Altmann 2005, s. 24–27).
2) **Matematická formulace hypotézy** – v tomto případě spolupracuje filolog s matematikem. Filolog musí matematikovi vysvětlit, čím je daný jev generovaný (tzn. za jakých okolností daný jev vzniká, v čem se projevují jeho vztahy k jiným jevům; Wimmer 2003, s. 15). Každá hypotéza ať už formulovaná slovně, nebo stanovena ve formě diferenciální rovnice, musí být transformována, aby mohla být testovaná statistickými metodami (Köhler a Altmann 2005, s. 25).
3) **Výběr dat** – filolog musí matematikovi poskytnout relevantní data (Wimmer 2003, s. 16)
4) **Testování** – úkolem fáze testování není poskytnout přímou odpověď a demonstrovat „pravdu“, ale vypočítat pravděpodobnost, která je podkladem pro naše rozhodnutí, jestli model jevu přijmeme nebo zavrhneme (Wimmer 2003, s. 16)
5) **Statistická interpretace** – výsledkem statistického testování je číslo, které statistik interpretuje, tj. zjišťuje pravděpodobnost chyby, které se dopouštíme, když přijímáme nevhodný model anebo když zamítáme vhodný model. Filolog musí sám rozhodnout, zda je jeho hypotéza přijatelná. Vhodnost modelu se dále induktivně ověřuje na co největším počtu případů a dále se začleňuje do existujících teorií, což jim zaručuje deduktivní potvrzení (Wimmer 2003, s. 16)
6) **Filologická interpretace** – výsledek testování se překládá zpět do řeči filologie, přičemž genezi, struktuře či chování se jevu se připisuje model, například křivka, rozdělení, proces, matematická struktura, která je odrazem filologické hypotézy (Wimmer 2003, s. 16).

1.1 Užití kvantitativních metod v lingvistice v Číně

Kvantitativně-lingvistické studium čínského jazyka nebylo po dlouhou dobu v popředí zájmu čínských badatelů, a i když se určité prvky kvantitativně-lingvistického výzkumu začaly objevovat již od 20. let 20. století, většinou se výzkum soustředil na statistické měření četnosti slov a matematické metody byli v lingvistice využívány zejména pro účely jazykových reforem, kompilace učebnic a sestavování korpusů (viz níže). Aplikace statistiky na výzkum jazyků sice položila základ pro další kvantitativní zkoumání čínštiny, není to však totéž jako samotný kvantitativně-lingvistický výzkum. Z toho důvodu existuje poměrně málo univerzálních jazykových zákonů formulovaných čínskými lingvisty. Větší oblibu si kvantitativní výzkum získal až v posledních letech. Zda jsou jazykové zákony a teorie, které byly objeveny zahraničními vědci, aplikovatelné i pro čínský psaný a mluvený jazyk, je nadále třeba zkoumat (Liu a Huang 2012, s. 181–182).

Dle Cornelie Schindelin (2008, s. 97–98) počátky moderního užití kvantitativních metod v lingvistice v Číně sahají již do počátku 20. let 20. století. Čínský pedagog Chen Heqin (Chén Hèqín 陈鹤琴; 1892–1982), který mimo jiné studoval v zahraničí, kde přišel do kontaktu s moderními vědeckými metodami, prováděl rozsáhlé studie zaměřené na zjištění četnosti znaků. Jeho práce byly určeny zejména pro pedagogické účely. Po vyhodnocení textových korpusů vytvořil seznamy znaků nejprve seřazených dle radikálů s uvedením absolutní frekvence v korpusu a poté seznam sestupně seřazený dle absolutní frekvence znaků, přičemž všechny znaky se stejnou frekvencí jsou uvedeny za příslušným číslem (Chen 1928). A právě četnost znaků byla ve středu zájmu učenců v Čínské lidové republice až do 80. let 20. století. V letech 1928 až 1988 bylo vytvořeno nejméně 29 seznamů často používaných znaků, které byly sestaveny různými způsoby a obsahovaly soupisy od přibližně 1 000 do přibližně 4 500 znaků, např. *Frekvenční slovník moderního čínského jazyka* (1986) (Xiàndài hànyǔ pínlǜ cídiǎn 现代汉语频率词典) obsahující 4 474 typů znaků a 31 159 typů slov (Schindelin 2008, s. 99, 101).

Další významnou osobností byl profesor ekonomie a lingvista se zaměřením na čínské znakové písmo Zhou Youguang (Zhōu Yǒuguāng 周有光; 1906–2017). Na začátku 80. let přišel s hypotézou sestupné účinnosti čínských znaků, která předpokládá, že nejčastějších 1 000 znaků pokrývá 90 % všech běžných textů a se znalostí 5 100 běžných znaků by člověk dokázal přečíst 99,99 % veškerého běžného textového materiálu

psaného v čínských znacích (Zhou 1980 a Zhou 1992). Jeho hypotéza byla pozdějšími výzkumy potvrzena (Schindelin 2008, s. 99–100).

Zjišťování frekvencí znaků a následné sestavování seznamů běžných moderních čínských znaků bylo opěrným bodem pro sestavování učebnic. Na konci 80. let bylo dostatečné množství materiálů ze studií frekvence, aby byl sestaven *Seznam běžně používaných znaků v moderní čínštině* (Xiàndài hànyǔ chángyòngzì biǎo 现代汉语常用字表) z roku 1988, který kompiloval práci Chen Heqina a další vzniklé frekvenční slovníky a seznamy. Seznam zahrnoval 3 500 znaků a po konzultaci s odborníky do něj byly zahrnuty i znaky, které se běžně vyskytují v mluveném jazyce a zřídka v jazyce psaném (Schindelin 2008, s. 104).

Neopomeňme zmínit, že kromě jazykové didaktiky pomohlo zjišťování frekvencí znaků i při výběru znaků při reformě písma v padesátých letech, která měla za cíl zvýšit gramotnost co nejširších vrstev obyvatelstva. V roce 1956 byl vyhlášen *Plán zjednodušení čínských znaků* (Hànzì Jiǎnhuà Fāng'àn 汉字简化方案), který měl zjednodušit písemnou soustavu třemi způsoby: zmenšením počtu tahů, zmenšením počtu znaků a zjednodušením způsobu psaní. Výsledkem bylo zjednodušení 515 čínských znaků a 54 prvků (Zádrapa 2009, s. 166–167). Kromě toho byl výzkum četností znaků a slov potřebný pro optimalizaci a standardizaci písma pro vydavatele, tiskárny a telegrafii.

Dalším významným lingvistou, který působí i v současné době, je Feng Zhiwei (Féng Zhìwěi 冯志伟; 1939–). Specializuje se na interdisciplinární výzkum v oblasti lingvistiky a počítačové vědy. Po studiích na Pekingské univerzitě a Čínské vědecko-technologické univerzitě působil střídavě na čínských a zahraničních univerzitách a ve vědeckých institucích. Na konci 70. let odjel do Francie, kde pod vedením profesora B. Vauquoise vytvořil první systém strojového překladu FAJRA z čínštiny do angličtiny, francouzštiny, němčiny, japonštiny a ruštiny. V 80. letech se dále zabýval strojovým překladem a počítačovou lingvistikou, ke konci 80. let jako hostující vědecký pracovník v Německu vytvořil první čínskou termínovou databázi GLOT-C. Byla to první čínská databáze termínů na světě používající čínské znaky. Zkoumal také míru entropie čínských slov. Profesor Feng má desetiletí zkušeností a poznatků v oblasti strojového překladu a je také představitelem oboru počítačové vědy s názvem NLP (Natural Language Processing, Zpracování přirozeného jazyka; Feng Zhiwei 2021).

V 90. letech byla pozornost soustředěna na řešení problémů, které se objevovaly v přechozích letech, zejména otázka segmentace slov a hledání principů, kterými by se

měla řídit konstrukce jazykových korpusů a databází, kterých v těchto letech začalo vznikat velké množství. Postupně se začaly vytvářet například dva velké korpusy, první z nich *Komentovaný korpus Lidového deníku* (Rénmín rìbào biāozhù yǔliàokù 人民日报标注语料库) se skládá z textů všech čísel oficiálního vládního deníku *Renmin ribao* z první poloviny roku 1998 a obsahuje přibližně 13 milionů znaků nebo přibližně 7,3 milionu slov. Práce na druhém z nich *Korpus moderního čínského jazyka* začala na počátku 90. let a na konci dekády korpus obsahoval 70 milionů znaků. Kompilace celého korpusu byla dokončena na konci roku 2001 s tím, že se plánovalo korpus rozšiřovat o 5 % ročně (Schindelin 2008, s. 105–107).

Jak již bylo zmíněno v úvodu, automatická segmentace slov čínských textů je velkým problémem, protože hranice slov nejsou v čínském textu nijak vyznačeny a jednotlivé znaky stojí vedle sebe, bez mezer (s výjimkou interpunkčních znamének). Čínská věta pak může být rozdělena na slova více smysluplnými způsoby. Prvním standardem ošetřující problematiku automatické segmentace byla doporučená *Norma pro segmentaci slov moderního čínského jazyka pro zpracování dat* (Xìnxī chǔlǐ yòng xiàndài Hànyǔ fēncí guīfàn 信息处理用现代汉语分词规范) GB/T 13715–92 z roku 1992, která obsahuje základní principy pro automatickou segmentaci slov. Cílem autorů bylo vytvořit pravidla, která zabrání nejednoznačnosti při segmentaci slov (Liu et al. 2013, s. 2). Při tvorbě pravidel vycházeli ze zkušeností získaných z korpusové analýzy frekvenčního slovníku moderních čínských slov, avšak ne všechna pravidla jsou řádně vysvětlena a v některých případech se dokonce nezdají být oprávněná (Schindelin 2008, s. 105–106). Slovo je zde například vágně definováno jako nejmenší jazyková jednotka, která může být použita samostatně (Zuìxiǎo de néng dúlì yùnyòng de yǔyán dānwèi. 最小的能独立运用的语言单位; GB/T 13715–92, s. 1). Tato norma a její aplikace jsou podrobně vysvětleny v Liu (1994).

Strojovým překladem, počítačovou lingvistikou a počítačovou terminologií se zabývá také Jie Chunyu (Jiē Chūnyǔ 揭春雨). Blíže se k tématu automatické segmentace a strojovém překladu v Číně budeme věnovat v kapitole 2.3.3.2. Další informace například viz Feng (1995) a Huang a Xue (2012).

S vývojem technologií bylo nutné řešit otázku vkládání znaků a slov do softwarových systémů. Pro usnadnění vkládání celých slov byla v roce 1995 přijata *Společná sada slov pro zadávání čínských znaků z klávesnice* (Hànzi jiànpán shūrù yòng tōngyòng cíyǔjí 汉字键盘输入用通用词语集) norma GB/T

15732, která obsahuje 43 540 běžných slov, včetně frází, z nichž nejdelší výraz je o délce 12 znaků. Při výběru slov a frází se pravděpodobně vycházelo z frekvenčního slovníku *Frekvenční slovník běžných moderních čínských slov* (Xiàndài Hànyǔ chángyòngcí cípín cídiǎn 现代汉语常用词词频词典) z roku 1990, který obsahuje celkem 77 482 slov (od jednoslabičných po sedmislabičné) a ve kterém je slovo vždy uvedeno ve znacích a v přepisu pinyin spolu s uvedením absolutní a relativní četnosti a kumulativní relativní četnosti (Schindelin 2008, s. 107). Zadávání pomocí pinyinu zůstává i nadále hlavní metodou psaní čínských znaků pomocí klávesnice. Více o kódování viz například (Zádrapa 2009, s. 89–91).

Pro metodu zadávání čínských znaků pomocí grafických prvků byl v roce 1997 přijat standard GF 3001-1997, který podporuje vývoj metod zadávání na základě grafických prvků. Sada má celkem 578 různých prvků a je založen na velmi rozsáhlé grafické analýze (Schindelin 2008, s. 108).

Ze současných čínských badatelů stojí za zmínku zejména Liu Haitao (Liú Hǎitāo 刘海涛; 1962–) profesor lingvistiky na Zhejiang University, který se se svým týmem aktivně zabývá kvantitativním a komplexním síťovým výzkumem jazyků (Complex network research). Liu Haitao v současné době působí jako člen redakčních rad v prestižních periodicích *Journal of Quantitative Linguistics* (Editorial board 2022) a *Glottometrics* (Glottometrics – About 2022) a je tak jedním z nejviditelnějších čínských kvantitativně zaměřených lingvistů. Liu Haitao společně s Feng Zhiweiem představili teorii pravděpodobnostního valenčního modelu pro zpracování přirozeného jazyka (Probabilistic Valency Pattern Theory for Natural Language Processing), která tradiční valenční teorii rozšiřuje o pravděpodobnostní prvek (Liu a Feng 2007). Liu Haitao publikoval v ČLR i zahraničí více než 60 prací s lingvistickou tematikou. Mnoho publikovaných prací vzniklo ve spolupráci s kolegou Heng Chenem (Chén Héng 陈衡).

Vedle Liu Haitaa zmiňme ještě Fan Fengxianga (Fàn Fèngxiáng 范凤祥; 1950–), který se zajímá o korpusovou a kvantitativní lingvistiku. Studoval kupříkladu náhodné pokrytí textové slovní zásoby (Random textual vocabulary coverage) v anglických textech. V roce 2008 byl jmenován pomocným redaktorem mezinárodně uznávaného lingvistického časopisu *Glottometrics*. Spolu s Gabrielem Altmannem a dalšími čínskými i evropskými jazykovědci publikoval řadu prací týkající se kvantitativně-lingvistického výzkumu.

Na Hong Kongské Polytechnické Universitě působí Chu-ren Huang (Huáng Jūrén 黃居仁) a jeho vědecká činnost se zaměřuje na oblast čínštiny a s tím spojenou počítačovou a korpusovou lingvistiku. Mimo jiné se zabývá také automatickým segmentováním čínských slov (Huang et al. 2007; Huang et al. 2008).

V neposlední řadě zmiňme profesorku Xinying Chen (Chén Xīnyíng 陈芯莹), jejíž výzkum na Fakultě zahraničních studií Xi'an Jiaotong University se zaměřuje na kvantitativní výzkum čínštiny a angličtiny a mezi její výzkumná témata patří komplexní jazykové sítě, teorie valence, korpusová lingvistika a další (Xinying Chen 陈芯莹 2022).

1.2 Užití kvantitativních metod v lingvistice na Taiwanu

Matematické lingvistice se věnují nejenom vědci pevninské Číny, ale o tento obor mají zájem i odborníci na Taiwanu. Konkrétně jmenujme například profesora Shu-Kai Hsieha (Xiè Shūkǎi 謝舒凱), který působí na National Taiwan University na pracovišti Graduate Institute of Linguistics a specializuje se na počítačovou a korpusovou lingvistiku, lexikální sémantiku, textovou analýzu, ontologii, lexikon a kognitivní neurovědu. Společně s výše uvedeným Chu-ren Huangem publikoval články ohledně automatického segmentování čínských slov (Huang et al. 2007; Huang et al. 2008). Další představitelkou je odborná asistentka Siaw-Fong Chung (Zhōng Xiǎofāng 鍾曉芳), která se na National Chengchi University věnuje výzkumu v oblasti korpusové lingvistiky a lexikální sémantiky.

1.3 Kvantitativní výzkum čínštiny mimo Čínskou lidovou republiku a Taiwan

Kvantitativním výzkumem čínštiny se začali vědci mimo Čínskou lidovou republiku a Taiwan intenzivněji zabývat od 90. let dvacátého století. První ucelenou studii o rozložení délek čínských slov na základě frekvenčních údajů přinesla Maria A. Breiter (1994), která poukázala, že délka lexémů v čínštině souvisí s dalšími důležitými rysy jako je frekvence, polysémie, slovní druh nebo jazykový styl. Z německý vědeckých pracovníků jmenujme například Hartmuta Bohna (1998), který se zabýval různými oblastmi kvantitativní lingvistiky aplikované na čínský jazyk a kvantitativně-lingvistické zákony se snažil také ověřovat (zejména platnost MALu na čínském znakovém písmu). Významným jazykovědcem, germanistou a obecným lingvistou je Karl-Heinz Best, který

studoval délky čínských slov a jejich frekvenci (Zhu a Best 1992; 1997; 1998). Další německý vědec Wolfgang Menzel zabývající se zpracováním přirozeného jazyka společně s čínskými kolegy publikoval práce ohledně čínských slov. V prvé řadě pomocí statistického výzkumu předkládají strategii pro odstraňování dvojznačnosti při segmentaci čínských slov (Qiao et al. 2008), dále navrhují nové hybridní schéma pro odhad frekvence čínských slov, které zároveň využívá více typů korpusů (Qiao et al. 2010).

Ze Spojených států amerických uveďme Eiji Nishimotov (2003), který měřil a porovnával produktivitu pěti čínských přípon 们 -men, 化 -hua, 儿 -r, 子 -zi a 头 -tou a na základě analýzy zjistil, že nejproduktivnější je přípona 们 -men a nejméně 头 -tou.

Kvantitativní lingvista Richard Sproat se ve spolupráci s kolegyní Chilin Shih v Bellových laboratořích zaměřovali na segmentaci čínských slov (Sproat et al. 1996) a pracovali na systémech převodu čínsky psaného textu na řeč (Shih a Sproat 1996).

Singapurský počítačový lingvista Kim Teng Lua publikoval práce týkající se sémantiky spojené s rozpoznáváním nejednoznačných čínských slov, rozlišování jejich významu a zjišťování hranic slov v textu (Gan et al. 1996; Lua 1995).

Le Quan-Ha a jeho kolegové se při působení na Queen's University Belfast zabývali Zipfovým zákonem pro *n*-gramy čínských a anglických slovních spojení a slov (Le et al. 2002; 2003; 2006). Zipfův zákon byl předmětem studia i u Shmuela Shtrikmana (1994), který zjistil, že zatímco rozdělení četností čínských znaků se výrazně odchyluje od Zipfova zákona, rozdělení četností čínských slov se Zipfovým zákonem řídí. Další vědec Ronald Rousseau společně s Zhang Qioaqiao (Rousseau a Zhang 1994) znovu analyzovali data George Zipfa o frekvenci čínských slov a zjistili, že jejich frekvence se řídí Zipfovým rozdělením.

Zájem o výzkum čínštiny z kvantitativního hlediska lze postupně pozorovat i českých kolegů, kteří i ve spolupráci s čínskými lingvisty zkoumají čínské texty. Jeden z výzkumů zaměřující se na sociolingvistický a psycholingvistický jev střídání kódů publikovali Wang Lin a Radek Čech (2016). Autoři zkoumali vliv střídání kódů (přechodu z čínského do anglického jazyka) na MAL. Vycházeli ze dvou druhů korpusů: čínsko-anglického korpusu střídání kódů a druhého výhradně čínsky psaného korpusu se zaměřením na jazykovou úroveň souvětí – klauze. Zjistili, že vztah mezi délkou souvětí a délkou klauze se obecně řídí MALem, avšak lepší výsledky vykazuje korpus psaný výhradně čínštinou.

Zájem o kvantitativní metody se probouzí i u českých studentů sinologie a mimo v úvodu zmíněných kolegyň z Univerzity Palackého začínají vznikat práce zabývající se tímto tématem i na dalších univerzitách. Například na Univerzitě Karlově vznikla bakalářská práce *Možnosti kvantitativního rozboru vybraných rysů současné čínštiny a čínských textů* (Mikulec 2017), ve které autor Petr Mikulec poskytuje základní přehled elementárních postupů a fungování kvantitativní lingvistiky na čínských textech.

1.4 Menzerath-Altmannův zákon

Existují různé kvantitativně lingvistické ekonomizující zákony zabývající se slovem, jako jsou například Zipfovy zákony nebo brevitův zákon (Brevity law; též nazývaný jako Zipfův zákon o zkracování). Tyto zákony však zkoumají jazykovou jednotku *slovo* izolovaně v rámci jedné jazykové úrovně, což pro naše záměry není dostačující. Abychom mohli ověřit naši hypotézu, je třeba zvolit prostředek, který zkoumá vztah více jazykových entit.

Právě na takový vztah upozornil v roce 1928 Paul Menzerath. Při rozboru německých slov zjistil, že *čím delší je slovo, tím kratší jsou průměrné délky jeho slabik*. Na jeho zjištění navázal až v roce 1980 Gabriel Altmann, který pro tuto vlastnost jazykového systému formuloval obecnější hypotézu: *čím delší je jazykový konstrukt, tím kratší jsou jeho konstituenty* (Altmann 1980, s. 124). Zavedením pojmů konstrukt a konstituent zobecnil Menzerathovu teorii a formuloval tak zákon, kterému se na počest obou vědců říká Menzerath-Altmannův zákon. Konstrukt je jazyková jednotka na vyšší jazykové úrovni a konstituent je jazyková jednotka na nejbližší nižší jazykové úrovni. G. Altmann věřil, že i na lingvistiku je třeba aplikovat metodologické principy a důsledně směřovat k vědecké explanaci (Hřebíček 2002, s. 11). Na základě hypotézy G. Altmann odvodil matematickou formuli, viz např. (Wimmer 2003, s. 104; Benešová 2011, s. 21):

> Předpokládejme, že x je délka konstruktu a y je délka konstituentu. Relativní změna délky konstituentů $\frac{dy}{y}$ je úměrná relativní změně délky konstruktu $\frac{dx}{x}$:
>
> $$\frac{dy}{y} \sim \frac{dx}{x}$$
>
> při dosazení koeficientu úměry $b > 0$ dostaneme diferenciální rovnici:
>
> $$\frac{dy}{y} = -b\frac{dx}{x}$$

Rovnici vyřešíme metodou separace proměnných a získáme:

$\ln|y| = -b \ln|x| + c,$ kde *c* je reálná konstanta.

Proměnné *x* a *y* jsou nezáporná čísla, můžeme odstranit absolutní hodnoty a substituovat parametr $A = e^c$, dostaneme tak jednoduchou variantu formule Menzerath-Altmannova zákona:

$$y = Ax^{-b}$$

kde *x* je délka konstruktu měřená v jeho konstituentech,

y je průměrná délka jeho konstituentů měřená v jednotkách na nejbližší nižší jazykové úrovni,

A, *b* jsou kladné reálné parametry.

Ze vzorce je patrné, že s rostoucí velikostí konstruktu klesá velikost konstituentu.

Profesor Jan Andres se svým týmem (2012) navrhuje čtyři matematické formule, kterými může být vyjádřen Menzerath-Altmannův zákon:

Model 1:	$y = y_1 x^{-b}$	kde $A = y(1)$ a $c = 0$
Model 2 (zkrácený vzorec):	$y = Ax^{-b}$	kde $c = 0$
Model 3:	$y = y_1 x^{-b} e^{c(x-1)}$	kde $Ae^c = y(1) = y_1$
Model 4 (kompletní vzorec):	$y = Ax^{-b} e^{cx}$	

kde *x* je délka konstruktu měřená v jeho konstituentech,

y je průměrná délka jeho konstituentů měřená v jednotkách na nejbližší nižší jazykové úrovni,

A, *b*, *c* jsou reálné parametry,

e = 2,718… je Eulerovo číslo.

Velikost parametru *A* řídí posun celé křivky nahoru nebo dolů vzhledem k ose *x*. Absolutní velikost záporného parametru *b* ovlivňuje strmost křivky. „Čím větší je záporná hodnota parametru *b*, tím prudčeji klesá křivka znázorňující funkci *y* (Hřebíček 2002, s. 55–56).“

I když je možné MAL vyjádřit pomocí více matematických vzorců, z nichž každý by měl odrážet zákon ve své heuristické podobě, z důvodu přehlednosti budeme nadále používat pouze zkrácenou verzi (Model 2). Použití této formule považují autoři článku *Optimization of Parameters in the Menzerath–Altmann Law, II* (Andres et al. 2014, s. 22) za vhodnou jednak pro vyšší, tak i pro nižší jazykové úrovně. Pro zájemce budou výsledky získané pomocí ostatních vzorců uvedeny pouze v příloze této práce.

Abychom tento vztah ilustrovali na konkrétním příkladě, představme si, že máme čtyřslabičné slovo **lingvistika**, délka konstruktu (slovo měřené ve slabikách) je $x = 4$ (tzn. čtyři slabiky ling-vis-ti-ka) a průměrná délka konstituentů (slabik měřených v grafémech) je $y = 2{,}75$ (tzn. 11 grafémů / 4 slabik = 2,75). Z tohoto příkladu je patrné, že konstrukt je zpravidla celé číslo, zatímco délka konstituentů bývá nejčastěji číslo desetinné.

Při analýze celého výběrového souboru na dané jazykové úrovni je třeba zohlednit všechny délky konstruktů x a v případě konstituentů je třeba sečíst všechny velikosti konstituentů, které tvoří konstituenty konstruktů o velikosti x a z tohoto součtu vypočítat průměrné hodnoty y. Konkrétně by tabulka pro výpočet hodnot mohla vypadat takto:

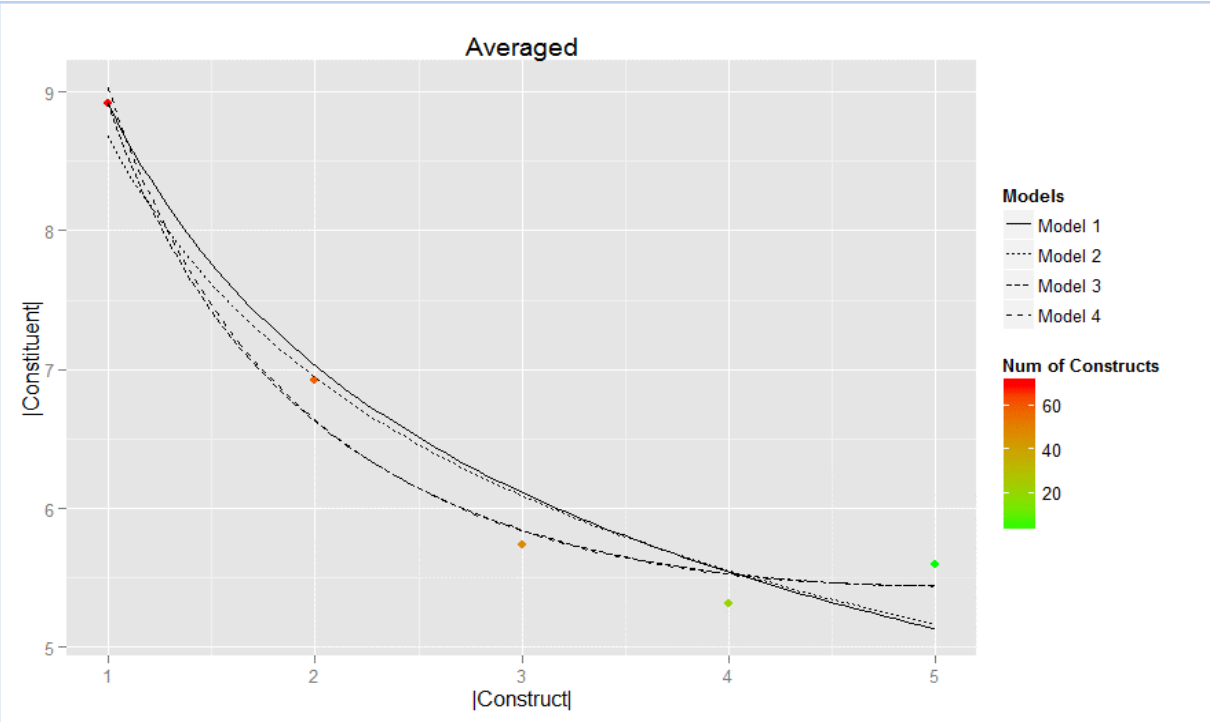
Tabulka 1 Příklad tabulky výpočtů MAL na jazykové úroveň U1: souvětí (měřené v klauzích) – klauze (měřená v průměrném počtu slov)

Délka souvětí v klauzích *x*	**Frekvence *z***	**Délka všech klauzí u jednotlivých souvětí ve slovech**	**Průměrná délka klauze ve slovech *y***
1	70	624	8,91
2	58	803	6,92
3	47	809	5,74
4	21	446	5,31
5	3	84	5,60

Zdroj: vlastní zpracování

Z pozorovaných hodnot y je zřejmé, že očekávaný pokles neplatí absolutně pro každou hodnotu zvlášť, ale projevuje se jako celková klesající tendence proměnné y spolu s rostoucím x (Hřebíček 2007, s. 84).

V grafickém vyjádření by tento vztah mezi dvěma veličinami měl klesající tendenci a mohl by vypadat následovně (viz Obrázek 1):



Obrázek 1 Grafické znázornění Menzerath-Altmannova zákona

Jsme si vědomi, že stejně jako u většiny jazykových jevů se mohou objevit výjimky, kdy pozorované tendence nejsou v souladu s MALem. Samozřejmě jsou i studie, které univerzální platnost MALu zpochybňují (Ferrer-i-Cancho et al. 2014) nebo jeho platnost nepotvrdily (Clink a Lau 2020). Navzdory tomu jsme se rozhodli pro ověření našeho výzkumu MAL použít, protože tento zákon umožňuje jazykovou jednotku *slovo* ověřit nejenom jako samostatně stojící jednotku, ale můžeme na ni nahlížet jako na jednotku, která je součástí celého řetězce. Nejenom, že MAL zkoumá vztah dvou jazykových entit, ale výzkum rozšiřuje na různé jazykové hladiny, na *slovo* tak můžeme nahlížet jako na konstrukt i konstituent a *slovo* lze ověřovat ve více rovinách.

V poslední době se testování MALu na čínském jazyce věnovali i čínští lingvisté, například již zmíněny Chu-ren Huang s kolegy zkoumali platnost MALu na čínských mluvených a psaných textech na jazykové úrovni souvětí – klauze a došli k závěru, že zákonitosti se projevují pouze ve formálních psaných textech (Hou et al. 2017).

Liu Haitao a jeho tým zkoumal platnost MALu na anglických textech různých funkčních stylů a ověřovali, zda hodnoty parametrů MALu indikují odlišnosti těchto funkčních stylů (Xu a He 2018). V následujícím výzkumu na čínských textech psaných

ve znacích se zabývali více jazykovými úrovněmi a zjistili, že při analýze funkčních stylů je třeba brát v úvahu rozdíly v hierarchických vztazích mezi jazykovými jednotkami na různých úrovních (Chen a Liu 2022).

Zákon inspiroval mnoho vědců a jeho platnost se snažili a stále intenzivně snaží ověřovat nejenom na různých jazycích (např. Torre et al. 2021), ale dokonce i v jiných vědách, jelikož analýzy lze univerzálně rozšířit na jakýkoli systém, který je tvořen hierarchií jednotek různých úrovní. Těmito oblastmi jsou např. evoluční biologie, genetika (Li, 2012; Matlach 2018; Matlach et al. 2022; Ferrer-I-Cancho a Forns 2009; Sun a Caetano-Anollés 2021), hudba (Boroda a Altmann 1991), sociologie a psychologie (Wang a Čech 2016), zákon byl také aplikován na znakový jazyk (Andres et al. 2019; Langer et al. 2020). Luděk Hřebíček upozornil na možnost aplikace tohoto zákona na teorii fraktálů a chaosu (Kułacka 2010, s. 257) a (Wimmer 2003, s. 103).

Cílem našeho experimentu je ověřit hypotézu, že ortografické slovo, které vznikne při převodu čínského textu do abecedy pinyin dle normy GB/T 16159–2012 jako část textu ohraničena mezerami, je v souladu s ekonomizujícími pravidly jazyka a shoduje se s obecně platnou definicí čínského slova. K ověření nám poslouží zkrácená verze Menzerath-Altmannova zákona (Model 2).

2 Metodologie

Experiment bude (s lehkými obměnami) proveden v krocích, podle kterých jsme postupovali v předchozím výzkumu, viz Motalová a Spáčilová (2013). Tyto kroky jsou následující:

1. Stanovení kritérií pro volbu výběrových souborů a jejich odůvodnění
2. Volba výběrových souborů
3. Stanovení a definování jazykových jednotek a úrovní
4. Segmentace a kvantifikace výběrových souborů
5. Testování spolehlivosti modelu pomocí statistických metod
6. Interpretace získaných dat

2.1 Kritéria pro volbu výběrových souborů

V ideálním případě bychom za výběrový soubor zvolili *soubor základní*, což je takový soubor, který obsahuje veškeré texty téhož typu vzniklé v jistém časovém období, např. veškeré texty generované v čínštině ve 20. století (Slavíčková 1988, s. 209). Z uvedeného příkladu je patrné, že získat všechny prvky této množiny je nemyslitelné, a i kdybychom takový soubor měli k dispozici, s ohledem na časovou náročnost je téměř nemožné celý základní soubor analyzovat. Proto je nutné přistoupit k analýze pouze části základního souboru, čímž získáme tzv. *výběrový soubor,* což je reprezentativní obraz základního souboru (Budíková et al. 2010, s. 13–14; Wimmer 2003, s. 21; Těšitelová 1987, s. 28). „Informace poskytované VS (výběrovým souborem) jsou adekvátnější ve smyslu větší orientace k účelu analýzy a jsou také spolehlivější, protože menší množství textů umožňuje zvýšit všestrannost popisu i důkladnost jeho kontroly (Slavíčková 1988, s. 209).“

Při volbě výběrových souborů je třeba vzít v potaz, jestli budeme zkoumat heterogenní či homogenní vzorky. Jak uvádí Sebastian Rasinger (2008), čím je základní soubor větší, tím je pravděpodobnější, že bude rozmanitější, a stejně tak i výběrové soubory, které ze základního souboru vycházejí, mohou být více heterogenní. Na jednu stranu mohou tyto výběrové soubory vykazovat stejné vlastnosti, které je řadí mezi členy základního souboru (např. všichni lidé mají společný znak lidství), ale zároveň se v různých atributech mohou odlišovat (každý člověk má jiné vlastnosti, věk apod.). Zato při zvolení menšího základního souboru je pravděpodobné, že samotný základní soubor

bude více homogenní a stejně tak i jeho části. Ani homogenita, ani heterogenita základního souboru není nutně špatná, ba naopak může být užitečná v závislosti na otázce výzkumu. Výběrový soubor by měl vždy být odrazem základního souboru, proto by se homogenita nebo heterogenita měla vždy odrážet i ve výběrových souborech. Pro vyvození závěrů o heterogenním základním souboru není možné použít homogenní vzorek a naopak (Rasinger 2008, s. 46–47). Pro většinu používaných statistických metod jsou však vhodné homogenní vzorky (Köhler a Altmann 2005, s. 28) a také my se pro náš výzkum budeme snažit volit homogenní vzorky.

Zdaleka ne všechny materiály jsou vhodné pro kvantitativní analýzu a jejich výběr podléhá různým kritériím. Již v roce 1987 Marie Těšitelová (1987, s. 19) definovala hlediska pro výběr vhodných vzorků, která lze univerzálně aplikovat i nyní. Podle ní existují hlediska kvalitativní a kvantitativní.

Kvalitativní hlediska (Těšitelová 1987, s. 19–25):

- jazyková – například čínský, český, slovenský jazyk a další
- psychologická – například výběr jazykového materiálu pro různou věkovou skupinu žáků
- sociologická (resp. sociolingvistická) – například analýza frekvence slov v odborných projevech mužů a žen
- tematická:
  - například frekvenční slovník zaměřený na společenské vědy nebo technický slovník apod.
  - texty zařazené dle stylu (například beletrie, tzv. věcná a užitková literatura, návodná literatura a tzv. apelativní literatura)
  - texty dle formy publikování (knihy, brožury, periodika, tiskopisy, plakáty)
- „sémiotická“ – například texty orientované na věc, texty orientované na 2. osobu, texty orientované na vlastní prožitek
- a jiná.

Kvantitativní hlediska (Těšitelová 1987, s. 28):

- systematický neboli mechanický výběr
- náhodný výběr
- výběr souvislých částí textu

Výběr kritérií pro volbu výběrových souborů je jedním z prvních důležitých kroků při kvantitativní analýze. Abychom však postupovali systematicky a mohli navázat na předchozí výzkum, v našem případě ponecháme kritéria pro volbu výběrových souborů stejná jako v minulých experimentech. Budeme zohledňovat jak kritéria kvalitativní, tak i kvantitativní. Těmito kritérii jsou jazykové kritérium (současná moderní čínština); literární žánry; aktuálnost; souvislost výběrových souborů a délka výběrových souborů. Jednotlivá kritéria rozebereme níže.

2.1.1 Současná moderní čínština

Experiment má sloužit k upřesnění definice jazykové jednotky slovo (cí 词) používané v současné čínštině. Proto se naše pozornost zaměří na výzkum současné moderní čínštiny, tj. putonghua (pǔtōnghuà 普通话), která vychází z pekingského dialektu. Konktrétně budeme zkoumat texty psané stylem baihua (báihuà 白话), který je v současnosti považován za standardní čínský psaný jazyk. Mluvená a psaná čínština se navzájem ovlivňují, proto níže nastíníme vývoj jak čínského psaného, tak i mluveného jazyka a pozornost budeme věnovat také fonetickým přepisům, které jsou pro náš výzkum klíčové.

Čínský jazyk se po dobu tří tisíc let vyvíjel ve dvou variantách – psané a mluvené podobě. Obecně platí, že **psaný** jazyk se v různé míře liší od mluveného ve všech jazycích. Čínština je však specifická v tom, že literární jazyk wenyan (wényán 文言), jehož gramatika a slovní zásoba vychází z klasické čínštiny z doby před dynastií Qin (Qín Cháo 秦朝, 221–206 BC) do období Wei-Jin (Wèi-Jìn 魏晉, přibližně 220–420), se v Číně v oficiálně uznávané literatuře používal více než dva tisíce let a podstata klasického *wenyanu* zůstávala po dobu několika staletí víceméně stále stejná, zejména co se týče jeho monosylabické povahy a gramatické struktury (oproti tomu lexikální složka jazykovým vývojem procházela) a v průběhu 6.–7. století se s mluvenou podobou čínštiny rozešel úplně (Chen 1993, s. 506–507; Chen 1999, s. 67–68; Vochala, Hrdličková 1985, s. 65–66). Od 7. století se vedle literárního jazyka *wenyan* začal formovat nový literární jazyk *baihua*, který byl založený na hovorovém jazyce. *Baihua* sloužil pro tzv. nižší literaturu, která nebyla oficiálně uznávaná, například pro výklady budhistických textů, povídky a hry (Vochala, Hrdličková 1985, s. 67; Chen 1993, s. 507). Postupem času však začala vznikat smíšená díla, která obsahovala jak literární jazyk *baihua*, tak i *wenyan*. Koncem 19. a počátkem 20. století byl literární jazyk *wenyan*

reformován a více se tak přiblížil hovorovému jazyku, jeho obliba se tak opět vrátila a používal se v společensko-politické a vědecké literatuře (Vochala, Hrdličková 1985, s. 67). Toto postavení si však udržel pouze do Májového hnutí (též Hnutí 4. května) v roce 1919, kdy došlo k literární revoluci, a za oficiální náhradu byl zvolen literární jazyk *baihua*, který více odpovídal mluvenému jazyku a byl tak lépe uchopitelný pro široké masy (Chen 1993, s. 510).

Následující desetiletí *baihua* postupně nahrazoval *wenyan* jako standardní psaný jazyk. „Formování norem pro moderní psanou čínštinu založenou na tradičním baihua bylo procesem vstřebávání rysů ze třech hlavních zdrojů: z ne-severočínských dialektů, klasické čínštiny a cizích jazyků[2]“ (Chen 1993, s. 510). *Baihua* se z těchto zdrojů postupně obohacoval v oblasti lexiky, gramatiky a stylistiky a ke konci 40. let 20. století se stal novým moderním literárním jazykem (Vochala, Hrdličková 1985, s. 68; Norman 1988, s. 247). Současná podoba *baihua* může být označována jako xiandai baihua (xiàndài báihuà 现代白话), což bychom mohli přeložit jako současná spisovná čínština (Sehnal 2006). *Baihua* je na poslech srozumitelná pro mluvčího toho dialektu, ve kterém je text předčítán nahlas. „Baihua ovlivňuje skrze čínské znaky slovní zásobu i gramatiku jednotlivých dialektů a stává se jejich organickou součástí, jednou ze stylistických vrstev“ (Sehnal 2006).

V současné době se literární jazyk *wenyan* používá pouze velmi omezeně. Hlubším studiem se zabývají např. studenti historie a literatury, ale využití v reálném životě je minimální (Norman 1988, s. 247). Kromě studijní oblasti se s ním můžeme setkat také při psaní úvodů k odborným studiím nebo v poezii (Kane 2009, s. 89). Další informace viz Chen 1999, s. 72, 76, 207–208 a Norman 1988, s. 136.

V případě, že mluvíme o písemném stylu, máme na mysli současnou podobu *baihua*, která je standardizovaná pro všechny Číňany bez ohledu na jejich rodný dialekt.

Stejně jako psaná čínština se i **mluvená** čínština v průběhu své více než tří tisícileté historie měnila a vyvíjela. Změnami procházela zvuková podoba čínského jazyka, měnila se jeho mluvnická stavba, funkční styly a k největším obměnám docházelo v případě slovní zásoby (Vochala, Hrdličková 1985, s. 48). Jelikož čínština byla a je zaznamenána formou znakového písma, které neodráží jeho zvukovou podobu, nebylo jednoduchým úkolem rekonstruovat fonetické změny v jejím historickém vývoji. Díky

[2] The formation of the norms for MWC [Modern Written Chinese] on the basis of the traditional baihua has been a process of absorbing features from three major sources: non-Northern Mandarin dialects, Classical Chinese, and foreign languages. Vlastní překlad autorky.

dochovaným písemným památkám a úsilí badatelů se však podařilo fonetickou výslovnost zrekonstruovat poměrně přesně (Švarný 1967, s. 10–11). Rekonstrukce fonologického vývoje čínštiny vycházela z čínských psaných záznamů jako jsou homofonní znaky, rýmy, transkripce (výslovnost cizích slov se zaznamenávala pomocí znaků a v případě, že známe původní cizí slovo, můžeme odhadnout jeho výslovnost) a dále z čínských výpůjček v japonštině, korejštině a vietnamštině (Baxter 1992, s. 11–14).

Abychom měli konkrétnější představu, co myslíme současnou čínštinou, uveďme příklad periodizace tohoto mluveného jazyka. Existuje více možných přístupů k periodizaci vývoje čínštiny. Některá členění jsou založená pouze na fonetické stránce jazyka, jako například periodizace známého švédského sinologa Bernharda Karlgrena (1889–1978) z let 1915–1926, která však již ve své původní podobě nebývá přijímána. Další možností, jak přistupovat k periodizaci, je využití gramatických změn (např. Jachontov) nebo se současně opírat o více kritérií, jako například významný lingvista Wang Li (Wáng Lì 王力; 1900–1986), který při členění vývojových fází čínského jazyka zohledňuje fonetické, lexikální i gramatické změny (Třísková 2010, s. 11; Švarný 1967, s. 11). Časové údaje i terminologie se u různých autorů liší, pro ilustraci uvádíme návrh členění čínštiny s přibližnou datací od Wang Liho (Wang Li 2004, s. 43–44)[3]:

Archaická čínština (Shànggǔqī 上古期)	asi od pol. 2. tisíciletí př. n. l. do 3.–4. stol. n. l.
Stará čínština (Zhōnggǔqī 中古期)	od 4. stol. do 12.–13. stol.
Nová čínština (Jìndài hànyǔ 近代汉语)	od 13. stol. do 19. stol.
Současná čínština (Xiàndài hànyǔ 现代汉语)	od Hnutí 4. května 1919 doposud

Po vzniku Čínské lidové republiky (ČLR) zahájila komunistická vláda jazykovou reformu a společný národní jazyk získal oficiální název *putonghua* (Třísková 2010, s. 14; Chen 1999, s. 23–24). Termín *putonghua*[4] se sice začal objevovat již na počátku 20. století (na sklonku dynastie Qing), ale současného významu nabyl až v roce 1955, kdy tento termín začal označovat současnou moderní čínštinu (Chen 1993, s. 508). *Putonghua*

[3] Překlady jsou uvedeny dle českého sinologa Oldřicha Švarného, viz (Švarný 1967, s. 11).

[4] Složka pǔtōng doslova znamená „běžný, obyčejný“, ve slově pǔtōnghuà má však význam „obecně rozšířený“ (Chen 1993, s. 508).

byla schválena jako společný jazyk ČLR v roce 1956, kdy také vznikla oficiální definice (Zhang 2005, s. 4–5):

普通话是以北京语音为标准音、以北方话为基础方言、以典范的现代白话文著作为语法规范。

> Pǔtōnghuà shì yǐ Běijīng yǔyīn wèi biāozhǔnyīn, yǐ běifānghuà wèi jīchǔ fāngyán, yǐ diǎnfàn de xiàndài báihuàwén zhùzuò wéi yǔfǎ guīfàn.

> „Putonghua je společný jazyk čínského národa, jehož fonetickým standardem je pekingská výslovnost, dialektálním základem jsou severní nářečí a mluvnickým standardem jsou vzorová díla napsaná současnou spisovnou čínštinou.“[5]

Díky tomu, že ve výslovnostní normě existuje jistá variabilita a definice *putonghua* není statická, jazyk se tak může přizpůsobovat změnám, které se neustále objevují. Již od roku 1955 se vláda ČLR snaží, aby byl jazyk *putonghua* rozšířen mezi veškeré obyvatelstvo, a to i v místech, kde se mluví místními dialekty. Jako prostředek šíření společného národního jazyka *putonghua* slouží škola, média, divadlo apod. (Wrenn 1975, s. 221, 225; Švarný 1967, s. 16).

Z výše uvedeného vyplývá, že **psaný a mluvený jazyk** se od sebe značně liší. Existují znaky jako například 此 (cǐ *toto*), přivlastňovací slovce 之 (zhī) apod., které se vyskytují v textu, ale v mluveném projevu se prakticky nepoužívají. Věty psaného jazyka bývají naopak komplikovanější a mohou se zde objevovat konstrukce klasické čínštiny, které jsou v mluvené čínštině vzácné (např. čtyřslabičná spojení). Proto je vždy třeba uvést, jakou konkrétní formu čínského jazyka zkoumáme. V našem případě se výzkum bude zabývat současnou čínštinou *putonghua* psanou stylem *xiandai baihua*.

2.1.2 Fonetické přepisy čínštiny

Čínské znakové písmo odráží zvukovou podobu jazyka pouze omezeně[6], proto lze v historii zaznamenat řady pokusů o zachycení výslovnosti čínských znaků. Snahy

[5] Překlad dle české sinoložky Hany Třískové (Třísková 2010, s. 14)

[6] Čínské znaky v sobě určité fonetické prvky nesou, např. fonogramy, které se skládají z významové složky (radikál) a fonetické složky (fonetikum), avšak fonetická složka nemusí vždy být spolehlivým ukazatelem výslovnosti.

o fonetický zápis čínštiny iniciovali jednak samotní Číňané, ale také cizinci. Vznikaly tak zápisy latinkové i nelatinkové, z nichž ty druhé obvykle vycházely z grafiky čínských znaků.

Abychom objasnili, proč jsme si za přepis čínštiny zvolili právě abecedu pinyin, uvedeme krátký historický vývoj fonetických přepisů a blíže se zastavíme právě u zvolené abecedy pinyin.

Za první fonetický zápis čínštiny může být považována metoda fanqie (fǎnqiè 反切) z období konce dynastie Východní Han (25–220 n. l.), která výslovnost neznámého znaku vymezila pomocí dvou znaků známých. Tato metoda byla v Číně používána po mnoho staletí (Třísková 2012, s. 24).

První výraznější potřeba zápisu čínských slov nastala v období jezuitských misií na počátku 17. století. Jezuitští misionáři začali jako první používat pro transkripci čínských znaků latinku (Třísková 1999b, s. 14). Problémem však byla rozdílná výslovnost jednotlivých znaků v různých oblastech Číny, lišila se především na území severu a jihu. Latinkové názvy tak měly různou podobu a bylo nesnadné je v různých přepisech identifikovat. Navíc začaly vznikat různé transkripce čínštiny, které byly založeny na fonetických a ortografických zvyklostech toho kterého jazyka. Tyto transkripce sice víceméně foneticky věrně zachycovaly čínskou výslovnost, byly však určeny pouze pro čtenáře v dotyčném jazyce a transkripce se navzájem velmi lišily (Palát 1999, s. 23).

Jednu z prvních řádných transkripcí čínštiny vytvořil francouzský jezuitský misionář Nicolas Trigault a v roce 1626 ji publikoval pod názvem *Pomůcka pro uši i oči západních učenců* (Xīrù ěrmùzī 西入耳目资). Nicolas Trigault přepracoval a rozšířil již vytvořený systém, který vypracoval jiný jezuitský misionář – Matteo Ricci. I když transkripce zaujala i čínské vzdělance, na čínské prostředí neměla téměř žádný dopad (Třísková 1999b, s. 14; Zádrapa 2009, s. 83; Mair 2002).

V čínském prostředí se o první fonetický zápis založený na modifikované latince zasloužil učitel a překladatel Lu Zhuangzhang (Lú Zhuàngzhāng 盧戇章; 1854–1928), který kvůli nízké gramotnosti v Číně navrhoval reformu v jazyce. V roce 1892 vydal zápis, který byl určen pro xiamenský dialekt a několik dalších čínských dialektů, pod názvem *Yimuliaoran chujie* (Yīmùliǎorán chūjiē 一目了然初阶). Později vytvořil fonetický zápis i pro pekingštinu: *Čínská fonetická abeceda* (Zhōngguó qièyīn zìmǔ 中国切音字母; Třísková 2012, s. 25; Wan 2014, s. 72 a Yimuliaoran chujie de

zhuyao neirong, Yimuliaoran chujie dao du 2019). Další informace viz Kaske (2008, s. 94–85).

Další významnou osobností byl Wang Zhao (Wáng Zhào 王照; 1859–1933), který v roce 1900 přišel s nelatinkovým přepisem *Mandarínská*[7] *abeceda* (Guānhuà héshēng zìmǔ 官话和声字母). Tento systém graficky vycházel z japonského slabičného písma katakana a byl založený na pekingském dialektu (Norman 1988, s. 258; Chen 1994, s. 367).

V 19. století se zahraniční sinologové snažili vytvořit jednotný systém přepisu čínského písma. Snaha však ztroskotala možná i z toho důvodu, že se sinologové snažili vytvořit systém transkripce založený na fonetice a ortografii svého jazyka a transkripce tak byly použitelné pouze v dotyčném jazyce. Proto se i nadále používaly jednotlivé národní transkripční systémy (Třísková 1999b, s. 13; Palát 1999, s. 23). Největší pozornost z těchto přepisů zasluhuje např. anglický přepis, označovaný jako *Wade-Gilesova transkripce*. Přepis vznikl ke konci 19. století, kdy Herbert Giles upravil již existující transkripci, kterou vytvořil sir Thomas Wade asi o čtvrt století dříve. V zahraničí to byl nejpoužívanější přepis čínštiny až do zavedení čínského latinizačního systému pinyin, a dokonce je možné se s ním setkat i dnes (Kane 2009, s. 22). Také profesor Jaroslav Průšek vycházel z transkripce Wade-Giles při vytváření prvního uceleného systému českého přepisu. Z dalších národních transkripcí lze uvést např. *Vissiérovu transkripci* pro frankofonní oblast, *Lessing-Othmerovu* a později *Behrsingovu transkripci* pro německou jazykovou oblast a *Palladijovu transkripci* pro rusky mluvící oblast (Třísková 1999b, s. 13).

Již od konce 19. století se v Číně objevovaly názory, že čínské znakové písmo je příliš složité a brání rozvoji gramotnosti a vzdělanosti, proto je třeba ho výrazně zjednodušit. V rámci Májového hnutí (1919) se většina reformátorů dokonce přikláněla k názoru zrušit čínské znakové písmo a nahradit ho latinkou (mezi nimi i např. spisovatel Lu Xun, Lǔ Xùn 鲁迅; Zádrapa 2009, s. 164).

V roce 1918 schválilo ministerstvo školství návrh abecedy *Zhuyin Zimu* (Zhùyīn Zìmǔ 注音字母), též známé jako *Bopomofo*, písmo pro zápis zvukové podoby znaků. Abeceda vycházela ze stavby čínské slabiky a z grafického hlediska reflektovala tahy

[7] Jednotný národní jazyk byl nejprve označován termínem guānhuà 官话, tzn. řeč mandarínů (úředníků), mandarínština (Třísková 1999b, s. 15)

čínských znaků. Od roku 1920 byla zaváděna do škol a v roce 1923 byl za výslovnostní normu přijat pekingský dialekt (Třísková 1999b, s. 17).

V letech 1925–1926 vznikl v Číně další přepis *Gwoyeu Romatzyh* (Guóyǔ Luómǎzì 国语罗马字) „latinka pro národní jazyk“, jehož autory byli představitelé hnutí za tzv. romanizaci národního jazyka. Autoři tuto přepisovou abecedu považovali za písmo rovnoprávné znakům a znaky podle nich měly být v budoucnu nahrazeny právě tímto přepisem. Avšak kvůli zbytečné složitosti abecedy jejich návrh neuspěl. Specifikem přepisu je včleňování tónů přímo do hláskového zápisu slabiky, nepoužívá tedy číslice nebo jiné speciální značky pro vyznačení tónu. Příkladem může být slabika guo: guō – **guo**, guó – **gwo**, guǒ – **guoo**, guò – **guoh** (Kane 2009, s. 23). Přesto je zrod této romanizační abecedy významným mezníkem, protože od té doby se v hnutí za reformu čínského písma stává hlavním proudem latinizace a při vytváření transkripcí se upouští od využití grafiky čínských znaků (Třísková 1999b, 17; Zádrapa 2009, s. 84).

Latinxua Sin Wenz (Lādīnghuà Xīn Wénzì 拉丁化新文字), neboli „nové latinizační písmo“, bylo druhou latinkovou transkripcí, která vznikla v letech 1929–1931 a za jejím vznikem stojí čínští a sovětští lingvisté, důležitou postavou byl zejména Qu Qiubai (Qú Qiūbái 瞿秋白). Písmo mělo sloužit pro šíření gramotnosti mezi čínskými dělníky žijícími v Sovětském svazu. Díky kampani proti negramotnosti dosáhlo po roce 1931 značného rozšíření i v Číně. Tento systém přepisování byl oproti předešlé romanizační abecedě podstatně jednodušší, ale jeho nedostatkem byla nižší vědeckost a přesnost. Nevýhodou bylo zejména neoznačování tónů a s tím související nemožnost rozlišit homofonní slabiky. Autoři této transkripce stejně jako autoři romanizační abecedy počítali s tím, že latinizační písmo později zcela nahradí čínské znaky. Avšak po založení Čínské lidové republiky v roce 1949 bylo jeho používání utlumeno a úkolem stranických a státních orgánů bylo vypracovat a rozšířit jednotnou spisovnou jazykovou normu (Chappell 1980, s. 105; Třísková 1999b, s. 17; Zádrapa 2009, s. 84).

V Číně se ujal až další vytvořený přepisový systém *Hanyu Pinyin* (Hànyǔ Pīnyīn 汉语拼音), který se používá dodnes. Viz podrobně níže: Abeceda Hanyu Pinyin.

Co se týče českého prostředí, první ucelený přepis čínštiny vytvořil prof. Jaroslav Průšek těsně před druhou světovou válkou. Tato transkripce však byla pro běžné čtenáře těžko uchopitelná, protože neodrážela přesnou výslovnost. Proto se koncem čtyřicátých a počátkem padesátých let ozývaly hlasy pro vytvoření nové transkripce čínských znaků. Vypracováním návrhu byl pověřen prof. Oldřich Švarný. Po diskusích byl v roce 1951

návrh přijat a vznikl tzv. standardní český přepis čínštiny (Palát 1999, s. 26–27). Tento přepis je pro běžného českého čtenáře lépe uchopitelný, protože staví na ortografických zvyklostech češtiny a čtenář tak dokáže poměrně přesně vystihnout výslovnost čínštiny. Standardní česká transkripce se dodnes používá kupříkladu v krásné literatuře, novinových textech a populárně naučné literatuře.

V následující tabulce uvádíme příklady přepisů čínských znaků pomocí několika vybraných transkripcí.

Tabulka 2 Příklady vybraných transkripcí

Zjednodušené čínské znaky	**Pinyin**	**Wade-Giles**	**Zhuyin Zimu**	**Gwoyeu Romatzyh**	**Česká transkripce**
中国	Zhōngguó	Chung^{1}-kuo^{2}	ㄓㄨㄥ ㄍㄨㄛˊ	Jonggwo	Čung1-kuo2
最近	zuìjìn	tsui^{4}-chin^{4}	ㄗㄨㄟˋ ㄐㄧㄣˋ	tzueyjinn	cuej4-ťin4
选材	xuǎncái	hsüan^{3}-ts'ai^{2}	ㄒㄩㄢˇ ㄘㄞˊ	sheuantsair	süan3-cchaj2
垂泣	chuíqì	ch'ui^{2}-ch'i^{4}	ㄔㄨㄟˊ ㄑㄧˋ	chweichih	čchuej2-čchi4
日杂	rìzá	jih^{4}-tsa^{2}	ㄖˋ ㄗㄚˊ	ryhtzar	ž'4-ca2

Zdroj: vlastní zpracování

2.1.2.1 Abeceda Hanyu Pinyin (Hànyǔ Pīnyīn Fāng'àn 汉语拼音方案)

Hanyu Pinyin Fang'an (zkráceně pinyin) je v současné době oficiální normou pro přepis výslovnosti čínských znaků do latinky a s jistou nadsázkou by se dalo říci, že se stal paralelním systémem písma. Latinkový systém pinyin vznikl v 50. letech 20. století v Číně jako výsledek práce lingvistů a primárně byl určen pro samotné Číňany k zachycení standardní výslovnosti znaků v *putonghua*. Prvotní koncept z roku 1955 využíval zkušeností předchozích latinizačních systémů vytvořených v Číně i v zahraničí a bral v potaz také nelatinkové systémy jako *Zhuyin zimu*. Po dobu dvou let procházel různými úpravami pod dohledem nejenom lingvistů, ale připomínkovali ho i další sektory společnosti, jako školství, věda, média, nakladatelství a zpravodajské kruhy, pošta,

železnice, zahraniční sinologové, cizinci studující čínštinu i široká veřejnost (Třísková 2010, s. 19).

Slovo pinyin se může používat též jako zkratka pro *Schéma čínské abecedy pinyin* (Hànyǔ Pīnyīn Fāng'àn 汉语拼音方案), což je oficiální standard ČLR, který byl ustanoven v únoru 1958. Tento dokument kodifikující abecedu pinyin zabírá pouhé čtyři strany a obsahuje pět oddílů: 1. Tabulka písmen, 2. Tabulka iniciál, 3. Tabulka finál, 4. Tónové značky, 5. Oddělovací znaménko (viz také kapitola 2.3.1 a 2.3.2). Od té doby je pinyin závaznou normou latinizaci čínského znakového písma. Oproti předešlým transkripcím čínštiny vytvořených v Číně se odlišuje v tom, že už od počátku nebyl určen pro nahrazení znaků, ale pouze k zaznamenání jejich výslovnosti (Chappell 1980, s. 105; Třísková 1999b, s. 19–20; Zádrapa 2006, s. 84, 92–94).

Pojem pīnyīn 拼音 doslova znamená „hláskovat, fonetizovat", nebo také „zachytit zvuky jazyka". Pinyin bývá některými považován za **abecedu** (zìmǔ 字母), někteří jej nazývají **transkripcí**, doslova „převedení zápisu" (zhuǎnxiě 转写), výjimečně je také označován za wénzì 文字, což znamená **písmo**.

Poslední zmíněné označení *písmo* však není přesné, protože pinyinu chybí podstatné atributy písma, např. ustálená a obecně vžitá ortografie nebo oficiálně potvrzený status písma. Další překážkou k uznání pinyinu jako *písmo* je jeho neschopnost přesně rozlišovat homofonní slabiky (Zhao a Baldauf 2008, s. 292). Ani označení *transkripce* neodpovídá povaze pinyinu. Ačkoliv byl pinyin vytvořen za účelem zachycení výslovnosti, nemůže být považován za fonetickou transkripci, protože jeden symbol nekoresponduje s jednou hláskou a nezachycuje skutečnou výslovnost (Třísková 2010, s. 22–23).

Stejně jako Hana Třísková (2010) se přikláníme k označení *abeceda pinyin*, jako určitý typ ortografického systému, který využívá grafémů latinské abecedy a funguje paralelně s grafickým systémem znakového písma, přičemž je vůči němu sekundární (Třísková 2010, s. 23).

Pinyin slouží zejména k šíření standardní čínštiny a její výslovnosti, k výuce čínštiny a také ke zvýšení gramotnosti v Číně. V běžném životě se využívá například při zpracování bibliografií, rejstříků, katalogů a slovníků, můžeme se s ním setkat na orientačních tabulích se jmény ulic, na nádražích, letištích nebo na vývěsních štítech obchodů. Díky pinyinu je vyhledávání v dokumentech mnohem snazší, protože slova mohou být seřazena abecedně podle výslovnosti. Systém pinyin se dále využívá pro

přepis čínských slov v jazycích, které používají latinku. Jeho funkce je neocenitelná také v moderních technologiích. Většina schémat pro fonetické zadávání používá k transliteraci čínských znaků právě pinyin (Zhao a Baldauf 2008, s. 11). Při vkládání znaků do počítače nebo mobilního telefonu se na klávesnici napíše výslovnost znaku pomocí pinyinu a z nabídky se následně vybere vhodný znak. Znalost zadávání pomocí pinyinu získává většina Číňanů ve škole (Zhao a Baldauf 2008, s. 122). Důležitou osobou, která stojí za vznikem hláskové abecedy pinyin, je čínský filolog, ekonom a člen výboru pro reformu čínského znakového písma Zhou Youguang, který je považován za „otce pinyinu" (Třísková 1999b, s. 19; Zádrapa 2006, s. 84, 92–94).

Pro praktické použití pinyinu bylo kromě *Schématu* z roku 1958 postupně vydáno několik dalších dokumentů. V roce 1974 vyšly zásady psaní čínských osobních jmen a mapa obsahující 3 800 zeměpisných názvů v pinyinu. O několik let později (1977) byl vydán atlas obsahující 18 000 zeměpisných názvů v přepisu pinyin. V témže roce byl na III. Konferenci o standardizaci zeměpisných názvů v Aténách přijat návrh, který určil systém pinyin jako mezinárodní standard pro přepis čínských zeměpisných jmen pro jazyky užívající latinku. Od roku 1978 bylo Státní radou ČLR rozhodnuto, že všechna osobní i místní jména vyskytující se v cizojazyčných latinku používajících textech publikovaných v Číně musí být přepisována v systému pinyin. V roce 1982 přijala latinkový systém pinyin i International Organization for Standardization (ISO) jako mezinárodní standard pro přepis *putonghua*, tzn. standardní čínštiny, do jazyků užívajících latinku. Blíže viz Vochala (1999, s. 35–36). Na Taiwanu byl Hanyu Pinyin zaveden od roku 2009 (Shih 2008).

Další dokument upřesňující pravopis této abecedy byl vydán v roce 1984 pod názvem *Základní pravopisná pravidla pinyinu (zkušení verze)* (Hànyǔ Pīnyīn zhèngzìfǎ jīběn guīzé (shìyòng gǎo) 汉语拼音正词法基本规则(试用稿), tento dokument upravil a sjednotil přepis znakové podoby slovníkových hesel do latinky. Také v dalších letech vycházely nové směrnice týkající se např. transkripce názvů čínských národnostních menšin, zaznamenávání titulů knih a časopisů, používání číslic v tištěných materiálech a používání interpunkčních znamének (Vochala 1999, s. 36).

Velmi obsáhle se tématu pinyinu a stavbě čínské slabiky věnuje Hana Třísková (2010).

2.1.2.2 Pravopisná pravidla pinyinu

Pinyin při svém vzniku nebyl považován za písmo, které by mělo nahradit čínské znaky, proto až do roku 1988 nebyla věnována pozornost jeho pravopisu. Je to pravděpodobně i z toho důvodu, že pinyin byl vytvořen zejména pro potřeby zachycení standardní výslovnosti znaků a *Hanyu Pinyin Fang'an* určoval, jakým způsobem se mají psát izolované slabiky, nikoli slova nebo souvislý text (Třísková 2010, s. 20). Teprve v roce 1988 vyšel dokument *Základní pravopisná pravidla pinyinu* (Hànyǔ Pīnyīn zhèngzìfǎ jīběn guīzé 汉语拼音正字法基本规则), ve kterém byly tyto problémy řešeny. V praxi se však bohužel stále setkáváme s nechutí Číňanů se těmito pravidly řídit a při přepisování znaků do latinky někdy píšou všechny slabiky oddělené mezerami, nebo naopak píšou slova bez mezer, nebo mezery umisťují libovolně bez ohledu na hranice slov (Třísková 2012, s. 20). V letech 1996 a 2012 vyšla pravidla v aktualizovaných podobách (viz podrobněji níže).

Na tomto místě je třeba připomenout, že předmětem této práce není porovnat, jak abeceda pinyin reflektuje výslovnost přepisovaných znaků nebo jakým způsobem se jednotlivé slabiky vyslovují. Práce se nezabývá fonetickou stránkou čínštiny, ale pouze ortografií pinyinu. Konkrétně se soustředí na čínská slova, která v textu psaném v čínských znacích nijak vyznačena nejsou, protože jednotlivé znaky k sobě těsně přiléhají, avšak v přepisu znaků do abecedy pinyin by měla být jednotlivá slova oddělena (neboli vytvořena) mezerami. Při určování hranic slov je pro nás zásadní dokument *Základní pravopisná pravidla pinyinu*[8] (Hànyǔ Pīnyīn zhèngcífǎ jīběn guīzé 汉语拼音正词法基本规则) (GB/T 16159–1996) a jeho novější modifikace *Základní pravopisná pravidla pinyinu*[9] (Hànyǔ Pīnyīn zhèngcífǎ jīběn guīzé 汉语拼音正词法基本规则) (GB/T 16159–2012), která dřívější dokument nahradila. Obě normy definují ortografická pravidla pro přepisování čínských znaků do abecedy pinyin a určují pravidla, kdy se mají slabiky psát dohromady a kdy odděleně.

Níže nastíníme, jaká pravidla jsou ve směrnicích obsažena a jaké změny byly provedeny v novější úpravě. První z těchto norem vyšla v roce 1996 a je rozdělená do čtyř částí, z nichž první tři jsou téma, terminologie a principy pro formulaci pravidel použitých v dokumentu. Čtvrtý nejobsáhlejší bod charakterizuje konkrétní pravidla pro přepis znaků do abecedy pinyin. Tato čtvrtá část má jedenáct podkapitol. První z nich

[8] V angličtině nese název *Basic rules of Hanyu Pinyin Orthography*

[9] V angličtině nese název *Basic rules of the Chinese phonetic alphabet orthography*

začíná obecnými ustanoveními. Druhá podkapitola se zabývá zápisem podstatných jmen a jsou zde zahrnuta i osobní a místní jména. Další podkapitoly se věnují zápisu sloves, adjektiv, zájmen, měrových jednotek, čísel, funkčních slov, čínských čtyřslabičných spojení chengyu (chéngyǔ 成语), psaní velkých písmen, pomlček a značení tónů.

Upravená verze normy z roku 2012 (*Basic rules of the Chinese phonetic alphabet orthography*) se od předchozí liší zejména svým rozšířením ze čtyř na sedm částí. Na první odlišnosti lze narazit již v úvodu, ve kterém jsou oproti předešlému dokumentu nejprve představeny oblasti, kterých se nový dokument týká, a dále jsou zde připojeny odkazy na další normy, které s tématem souvisí. Mimo jiné autoři pracují s odlišnou definicí čínského slova.

Definice slova z roku 1996 se objevuje v kapitole 3: Principy pro formulování pravidel (Zhìdìng yuánzé 制定原则) a v kapitole 4.1 Obecné pokyny (Zǒng yuánzé 总原则):

3.1 以词为拼写单位，并话当考虑语言，语义等因素，同时考虑词形长短适度。

Yǐ cí wèi pīnxiě dānwèi, bìng huà dāng kǎolǜ yǔyán, yǔyì děng yīnsù, tóngshí kǎolǜ cíxíng chángduǎn shìdù.

„Aby slovo bylo základní pravopisnou jednotkou, je třeba vzít v úvahu fonologické, sémantické a další faktory, a zároveň zohlednit délku slova.“

4.1.1 拼写普通话基本上以词为书写单位。

Pīnxiě Pǔtōnghuà jīběnshang yǐ cí wèi shūxiě dānwèi.

„V latinizovaném přepisu moderní standardní čínštiny, tj. putonghua, je slovo v podstatě považováno za pravopisnou jednotku.“

Definice slova v normě z roku 2012 je uvedena v části 3: Terminologie a definice (Shùyǔ hé dìngyì 术语和定义):

3.1 词 *cí* slovo

语言里最小的，可以独立运用的单位。

Yǔyánlǐ zuìxiǎo de, kěyǐ dúlì yùnyòng de dānwèi.

„… je nejmenší jazyková jednotka, která může být použita samostatně.“

Jazykovou jednotkou slovo se budeme blíže zabývat v kapitole 2.3.3.

Samotná pravidla pro přepis do abecedy pinyin začínají od bodu pět. Oproti starší verzi jsou části jednak jinak přeskupené a jednak se v některých bodech liší i obsahově. Týká se to např. bodu 5.4 (stará norma bod 4.1.4), ve kterém je nově uvedeno, že při zaznamenání reduplikace AABB není třeba vkládat pomlčku. Rovněž v případě některých čtyřslabičných výrazů chengyu (chéngyǔ 成语), které z důvodu rytmiky nelze rozdělit, je pomlčka vypuštěna (bod 6.1.12.1).

Další změny lze najít v bodu 6.2.3 (stará norma bod 4.2.5). Původem nečínská jména osob a míst se podle předešlé normy měla psát podle originálního znění (např. Ulanhu 乌兰夫; Marx 马克思 ad.). Novější norma však uvádí, že tato jména mají být zapisována přepisem pinyin (tzn. Wūlánfū 乌兰夫; Mǎkèsī 马克思).

Bod 6.1.5.3 v nové verzi normy se týká číslovek a konkretizuje zápis číslovky 10 (shí 十) v řádech (např. deset stomiliónů 十亿 lze zapsat zvlášť shí yì nebo dohromady shíyì), zatímco stará norma tuto číslovku blíže nespecifikuje. V bodu 6.1.9.1 se objevuje variantní zápis atributivního slovce de 的. Nově je možné v některých případech volit mezi připojením nebo nepřipojením atributivního slovce k adjektivu (např. přivlastňovací zájmeno *moje* 我的 může být zapsáno dohromady jako wǒde, nebo také zvlášť jako wǒ de). Kromě toho se v tomto bodě objevují i další případy, kdy je dle nové normy možné přepisovat znak de 的 dohromady společně s předchozí slabikou. Variantní zápis se týká i znaku de 得, příklady viz Tabulka 3. Dále stará norma v některých případech používá spojovník pro onomatopoia zatímco nová verze spojovníky vypouští (bod 6.1.11).

V nové verzi je také přidána celá nová část číslo sedm. Zabývá se dalšími oblastmi přepisu pinyin, které v minulé verzi normy nebyly zmíněny. Avšak v praxi se některá pravidla používala již od roku 1996 (např. bod 7.1). Ale body 7.3 a 7.4 se zabývají zcela novými oblastmi. Podle bodu 7.3 je možné zdůraznit neutrální tón tím, že se za danou slabiku vloží tečka. Podle bodu 7.4 je možné zapsat dohromady víceslabičné struktury, které představují kompletní strukturu. Výše uvedené změny jsou pro přehlednost uvedeny v Tabulce 3.

Tabulka 3 Srovnání norem z roku 1996 a 2012

Čínské znaky	**Převod dle normy 1996**	**Převod dle normy 2012**
来来往往	láilai-wǎngwǎng	láiláiwǎngwǎng
说说笑笑	shuōshuo-xiàoxiào	shuōshuōxiàoxiào
乌兰夫	Ulanhu	Wūlánfū
阿沛阿旺晋美	Ngapoi Ngawang Jigme	Āpèi Āwàngjìnměi
伦敦	London	Lúndūn
马克思	Marx	Mǎkèsī
东京	Tokyo	Dōngjīng
十亿零七万二千三百五十六	**shí yì** líng qīwàn èrqiān sānbǎi wǔshíliù	**shí yì** líng qīwàn èrqiān sānbǎi wǔshíliù / **shíyì** líng qīwàn èrqiān sānbǎi wǔshíliù
我的	wǒ de	wǒde / wǒ de
商店里摆满了吃的, 穿的, 用的。	Shāngdiàn li bǎimǎnle chī de, chuān de, yòng de.	Shāngdiàn li bǎimǎnle chī de, chuān de, yòng de. / Shāngdiàn li bǎimǎnle **chīde, chuānde, yòngde.**
红得很	hóng de hěn	hóng de hěn / hóngde hěn
写得不好	xiě de bù hǎo	xiě de bù hǎo / xiěde bù hǎo
叽叽喳喳	jiji-zhazha	jījīzhāzhā
大公鸡喔喔啼。	Dà gōngjī wo-wo-tí.	Dà gōngjī wōwō tí.
层出不穷	céngchū-bùqióng	céngchūbùqióng

Zdroj: Vlastní zpracování

V případě znaků de 的 a de 得, u kterých novější norma umožňovala dva způsoby zápisu, jsme volili oddělený zápis. Stažený zápis jsme volili pouze u případů, které byly v normě konkretizovány.

Kromě výše uvedené směrnice (upravená verze z roku 2012) nám jako zdroj pro segmentaci slouží kniha *Chinese Romanization: Pronunciation and Orthography* od autorů Yin Binyong and Mary Felley (1990). Autoři na téměř 600 stránkách podrobně vysvětlují ortografická pravidla a ilustrují je na příkladech. Po krátkém úvodu do fonetiky čínského jazyka kniha postupně přibližuje pravidla pro jednotlivé slovní druhy, idiomy a větné struktury. V závěru knihy jsou uvedeny praktické příklady textů převedených dle pravidel do abecedy pinyin. Autoři tak ukazují aplikovatelnost pinyinu na různé literární žánry.

V neposlední řadě je třeba zmínit, že čínština je jazyk tónový a moderní standardní čínština má tóny čtyři. Tóny značí určitý melodický průběh slabiky a rozlišují lexikální význam morfémů nebo slov. Pokud bychom zapsali čínskou slabiku přepisem pinyin, tón by byl značen diakritickou značkou nad hlavní samohláskou slabiky, délka slabiky v grafémech by se tedy nezměnila. V případě segmentace textů tedy tón slabiky nebude zohledněn.

Oproti češtině čínština nerozlišuje krátké a dlouhé samohlásky a délka vyslovené samohlásky tak nerozlišuje lexikální význam, ale vyjadřuje zdůraznění.

2.1.2.3 Zpracování textových souborů

V rámci naší práce budou za výběrové soubory vybrány texty psané v čínských znacích, které budou následně převedeny do abecedy pinyin podle pravidel uvedených v normě *Basic rules of the Chinese phonetic alphabet orthography* (GB/T 16159–2012). Kromě textů psaných v čínských znacích budou zvoleny texty, které jsou napsané čínskými autory přímo v abecedě pinyin a ty nebudeme nijak převádět či upravovat.

Abychom lépe ilustrovali, jakým způsobem jsme text z čínských znaků převáděli do pinyinu, uvádíme konkrétní příklad vybrané části textu z Výběrového souboru 1: Yu Hua – *Kamarádi* (viz kapitola 2.2.1). Původní text ve znacích jsme nejprve automaticky převedli do pinyinu pomocí softwaru Wenlin 4, funkce *Make transformed copy – Pinyin transcription* a následně jsme nabízené převody zkontrolovali a opravili dle normy GB/T 16159–2012.

Původní text:

大名鼎鼎的昆山走出了家门，他一只手捏着牙签剔牙，另一只手提着一把亮晃晃的菜刀。他扬言要把石刚宰了，他说：就算不取他的性命，也得割下一块带血的肉。至于这肉来自哪个部位，昆山认为取决于石刚的躲闪本领。

Text převedený pomocí Wenlinu 4 (funkce *Make transformed copy – Pinyin transcription*):

Dàmíngdǐngdǐng de kūn shān zǒuchūle jiāmén, tā yī zhīshǒu niēzhe yáqiān tīyá, lìngyī zhǐ shǒutízhe yī bǎ liànghuǎnghuǎng de càidāo. Tā yángyán yào bǎ shí gāng zǎi le, tā shuō: jiùsuàn bù qǔ tā de xìngmìng, yěděi gē xià yīkuài dài xuè de ròu. Zhìyú zhè ròu láizì nǎge bùwèi, kūn shān rènwéi qǔjué yú shí gāng de duǒshǎn běnlǐng.

Ručně opravený text dle normy GB/T 16159–2012 (změny oproti automatické segmentaci jsou vyznačeny červeným tučným písmem):

Dàmíngdǐngdǐng de **K**ūn **S**hān zǒuchūle jiāmén, tā yī zhī_shǒu niēzhe yáqiān tīyá, lìng_yī zhǐ shǒu_tízhe yī bǎ liànghuǎnghuǎng de càidāo. Tā yángyán yào bǎ **S**hí **G**āng zǎi le, tā shuō: jiùsuàn bù qǔ tā de xìngmìng, yě_děi gē xià yī_kuài dài xuè de ròu. Zhìyú zhè ròu láizì nǎge bùwèi, **K**ūn **S**hān rènwéi qǔjué yú **S**hí **G**āng de duǒshǎn běnlǐng.

V průběhu segmentace všech výběrových souborů, které jsme převáděli do pinyinu, jsme se setkali se slovy, jejichž zápis byl nejednoznačný. V těchto nejasných případech jsme se vždy snažili znaky přepisovat podle normy GB/T 16159–2012 a pokud tyto případy nebyly v normě definované, opírali jsme se o knihu *Chinese Romanization: Pronunciation and Orthography*, která obsahuje velké množství ortografických příkladů. Jsme si vědomi, že tato kniha vyšla již v roce 1990, proto bylo nutné opírat se pouze o ty kapitoly, které jsou v souladu s novou normou. Konkrétní příklady sporných segmentací na slova uvedeme v kapitole 2.3.3 Slovo (cí 词).

2.1.3 Literární žánry

Naše práce se zabývá zkoumáním různých výběrových souborů, a proto považujeme za nutné uvést, z jakých literárních žánrů budeme vybírat. Vzhledem k tomu, že již při vymezení samotného termínu *literární žánr* panuje jistá nejednoznačnost, nejprve uvedeme, co pojmem literární žánr myslíme. Při vymezení tohoto pojmu se budeme opírat o definici literárního teoretika Eduarda Petrů, která zní následovně „Jako literární žánr označujeme ty literární útvary, které se realizují uvnitř literárních druhů (epos, komedie, hymnus apod.), s vědomím, že tyto literární žánry jsou dále vnitřně diferencovány na žánrové varianty (například milostný román, historický román, sociální román atd.) a využívají různých *žánrových forem*" (Petrů 2000, s. 71).

V nejobecnější rovině, tedy rovině nadřazené literárním žánrům, Eduard Petrů (2000) rozděluje tři literární druhy: *lyriku, epiku* a *drama*. My se zaměříme pouze na *epiku*, tj. díla rozvíjející příběh ve formě vyprávění. Mezi epické literární žánry můžeme zařadit např. epos, román, legendu, povídku, bajku, reportáž, cestopis atd. Epická díla se vyznačují syntaktickým propracováním, logicky uspořádaným slovosledem a oproti mluvené podobě jazyka se v nich nevyskytuje tolik opakovaných a redundantních výrazů.

Předmětem této práce není komparace žánrů, proto není nutné, aby byl každý výběrový soubor jiného žánrového zařazení. Z důvodu rozsahu se nám jako nejvhodnější jeví zaměřit se v bádání na kratší literární útvary – v našem případě to budou zejména povídky. Kromě nich výběr doplníme o kratší text z učebnice (důvod volby viz kapitola 2.1.6).

2.1.4 Aktuálnost

Předkládaná práce se zabývá výzkumem současné moderní psané čínštiny, proto je třeba brát ohled i na aktuálnost výběrových souborů. Výběrové soubory musí být psané stylem *baihua*, který je prosazován od Májového hnutí 1919. Jazyk se samozřejmě neustále vyvíjí a *baihua* používaná v roce 1920 se od současné moderní psané čínštiny výrazně liší zejména v oblasti gramatika a slovní zásoby (Chen 1999, s. 82), proto se s ohledem na aktuálnost snažíme přiblížit co nejvíce současnosti a vybrat texty nynějších autorů. Dále jsme chtěli navázat na předchozí výzkum a využít jeden z předchozích výběrových souborů, který tentokrát budeme segmentovat novým způsobem. Proto jsme časové rozmezí vymezili na texty vydané od roku 1995.

2.1.5 Souvislost výběrových souborů

Ve spojitosti s aplikovanými metodami mohou být předmětem kvantitativně-lingvistického výzkumu jak texty souvislé, tak i nesouvislé. V našem případě se nám jeví použití souvislých útvarů jako nejvhodnější volba pro námi zvolený typ zkoumání. Stejně jako Luděk Hřebíček (2002, s. 10) nechápeme text jako množinu vět, ale jako propojený celek skládající se ze všech obsažených jednotek (hlásek, slabik, morfémů, slov a vět), mluvíme tedy o souvislých útvarech (Hřebíček 2002, s. 43; Hřebíček 2007, s. 27).

Dále se držíme pravidla, že je třeba vyhodnotit celý text jako celek, a ne pouze jeho část nebo náhodné výběry slov. Pokud se jedná o delší úseky, je možné texty rozdělit na kompaktní části (např. kapitoly) a vyhodnotit je samostatně. Zkoumané části by však měly být homogenní, protože pouze v homogenních textech je možné udržet konstantní proporce (Altmann a Meyer 2005, s. 45). Další podmínkou je, že mezi sebou nemůžeme míchat více různých textů, byť souvislých (Wimmer 2003, s. 21, 89).

Z důvodu udržování koherence textu musíme vyloučit některé literární žánry, které z našeho pohledu jen stěží můžeme považovat za souvislé texty. Pro náš výzkum se například z důvodu rozsáhlé délky nehodí celý román, vhodné nejsou ani celé sbírky povídek či seznamy tvořené jednoslovnými nebo několikaslovnými výrazy bez gramatické souvislosti.

2.1.6 Délka výběrových souborů

Při volbě délky výběrových souborů je třeba mít na zřeteli, aby bylo získáno dostatečné množství vstupních dat. Výběrové soubory nesmí být ani příliš krátké, ani příliš dlouhé, a to z toho důvodu, aby se projevily vzájemné vztahy jazykových jednotek.

V minulých experimentech jsme pracovali s výběrovými soubory, jejichž délka se pohybovala mezi 2 500 – 5 500 znaků (bez mezer). Tentokrát jsme se rozhodli, že délku výběrových souborů ještě navýšíme, abychom dostali co nejvíce relevantních dat. S ohledem na časovou náročnost segmentací jsme horní hranici délky textů zvolili na 10 000 znaků.

Již při selekci vhodných výběrových souborů jsme zamýšleli porovnávat nejenom texty psané čínským znakovým písmem následně převedené do latinky, ale mezi výběr jsme plánovali zařadit také texty psané přímo v latince, tj. v pinyinu. Takovéto texty, které by byly psané čínskými autory, však vznikají poměrně zřídka, protože pro čínské autory je přirozené používat čínské znakové písmo. Proto jsme se rozhodli kromě povídek zařadit mezi výběrové soubory i učebnicové texty, které jsou pro potřeby studentů psané

přímo v pinyinu a navíc procházejí korekturou rodilých mluvčích. Nevýhodou učebnicových textů psaných v pinyinu zpravidla je, že jsou svou povahou kratší délky, protože vznikají zejména pro potřeby studentů začátečníků či mírně pokročilých (delší texty jsou již psány pouze čínskými znaky). Námi zvolený soubor učebnicových textů má sice délku pouhých 461 slov, což by odpovídalo 605 znakům (bez mezer; viz kapitola 2.2.5), ale i přes svou kratší délku ho zařadíme mezi náš výběr, abychom mohli zjistit, jestli zápis slov v pinyinu (tedy oddělování slov mezerami) odpovídá normě GB/T 16159–2012. Případně zjistíme, v jakých případech se segmentace liší od této normy, a budeme ověřovat, jestli pravidla lépe odpovídají ekonomizujícím zákonům, což ověříme kvantitativně-lingvistickými metodami.

Kromě učebnicových textů se nám nakonec podařilo získat ještě další výběrový soubor psaný formou abecedy pinyin. Jedná se o deník čínské autorky Zhang Liqing (`Zhāng Lìqīng`), který disponuje dostatečnou délkou 2 771 slov (viz kapitola 2.2.4). I tento text bude předmětem naší analýzy z důvodů uvedených výše.

Rozmezí délek výběrových souborů se tedy bude pohybovat mezi 600 – 10 000 znaků (bez mezer). Například u nejdelšího výběrového souboru od Yu Hua *Vítězství ženy* odpovídá 9 658 znaků 5 786 slovům.

2.2 Výběrové soubory

Jak již bylo naznačeno výše, naším záměrem bylo analyzovat nejenom texty převedené z čínských znaků do abecedy pinyin dle normy GB/T 16159–2012, ale mezi výběrové soubory jsme plánovali zařadit i texty psané přímo v pinyinu. Cílem je zjistit, jestli texty psané přímo v pinyinu dodržují pravidla definovaná touto normou. V případě, že ne, budeme pomocí kvantitativně-lingvistických nástrojů zjišťovat, jestli použité členění na ortografická slova více odpovídá ekonomizujícím zákonům.

Na základě výše uvedených kritérií jsme zvolili pět výběrových souborů: tři povídky a jeden učebnicový text. Při vyhledávání vhodných výběrových souborů jsme narazili i na deník, který byl zaznamenán čínskou autorkou přímo v abecedě pinyin, který jsme také zařadili mezi zkoumané soubory.

Autorem prvních dvou povídek *Kamarádi* a *Vítězství ženy* je čínský spisovatel Yu Hua. Další povídku *Život, jak mu rozumím* sepsal čínský autor Han Han, následující výběrový soubor originálně psaný v abecedě pinyin *Deníkové záznamy v piyinu,* kapitola

Mihuo pochází od autorky Zhang Liqing a posledním výběrovým souborem jsou učebnicové texty z publikace *Integrated Chinese Level 1 Part 2*. Tyto zvolené výběrové soubory jsou blíže specifikovány níže.

2.2.1 Yu Hua (Yú Huá 余华): *Kamarádi* (Péngyou 朋友)

První dva výběrové soubory pocházejí z pera známého čínského autora Yu Hua, který se narodil roku 1960 v Hangzhou v čínské provincii Zhejiang. I když vystudoval medicínu a zpočátku pracoval jako zubař, toto povolání ho nenaplňovalo a rozhodl se pro literární dráhu. Svou první tvorbu začal publikovat ve svých 25 letech. Je držitelem několika mezinárodních literárních ocenění a jeho díla jsou překládána do mnoha světových jazyků (Zhao 1991, s. 415–420). Vzhledem k jeho čínskému původu a vlivu na širokou veřejnost můžou být jeho texty považovány za vhodné výběrové soubory.

První povídka *Kamarádi* byla zveřejněna v roce 1998 a nadále na ni v textu budeme odkazovat jako Výběrový soubor 1. Je psaná ve zjednodušených znacích a obsahuje 7 527 znaků (bez mezer). Výběrový soubor může být považován za souvislý text, protože neobsahuje žádné obrázky ani grafy s popisky. Text je členěn do odstavců, které vždy začínají na novém řádku a jsou odskočeny od okraje. Výběrový soubor je strukturován srozumitelně, např. úseky textu s přímou řečí jsou vyznačeny uvozovkami. Povídka je psaná pouze čínskými zjednodušenými znaky a nevyskytují se zde arabské číslice ani latinka. Kromě čínských znaků autor operuje s devíti různými interpunkčními znaménky (viz Tabulka 4) a jejich použití je standardní. Pouze u použití uvozovek jsme se setkali s případy, kdy uvozovky neoddělovaly klauzi (podrobně vysvětlíme níže).

Tabulka 4 Použitá interpunkční znaménka

Český název	Interpunkční znaménko	Pinyin	Znaky
Čárka	,	dòuhào	逗号
Tečka	。	jùhào	句号
Otazník	?	wènhào	问号
Středník	;	fēnhào	分号
Uvozovky	“ ”	yǐnhào	引号
Elipsa	……	shěnglüèhào	省略号

Dvojtečka	：	màohào	冒号
Čárka pro výčet	、	dùnhào	顿号
Dlouhá pomlčka	——	pòzhéhào	破折号

Zdroj: čínská terminologie převzata z (Biaodian fuhao, © 2017); vlastní zpracování

Pro účely experimentu jsme čínské zjednodušené znaky převedli do abecedy pinyin dle normy GB/T 16159–2012. Po převedení textu do latinky text obsahuje 4 571 slov.

2.2.2 Yu Hua (Yú Huá 余华): *Vítězství ženy* (Nǚrén de shèngli 女人的胜利)

Stejně jako první výběrový soubor je i tato povídka od autora Yu Hua psaná v zjednodušených znacích a originální text má 9 658 znaků (bez mezer). Nadále ji budeme značit jako Výběrový soubor 2. Povídka byla dokončena v roce 1995. Text je rozdělen do šesti částí (podkapitol) a každá z částí je označena číslem jedna až šest. Pro účely segmentace jsme však toto členění nebrali v potaz, protože výzkum je zaměřen pouze na zkoumání souvětí jako nejvyšší jazykové jednotky. Jelikož v tomto případě text na sebe významově navazuje (členění textu tedy neovlivňuje souvislost textu) a navíc se zde nevyskytly žádné grafy ani obrázky s popisky, může být výběrový soubor považován za souvislý text. Odstavce jsou stejně jako u předešlé povídky v textu jasně vyznačeny odsazením od okraje.

Kromě čínských znaků se zde několikrát vyskytla telefonní čísla zapsána pomocí arabských číslic, v segmentaci jsme je však nezohlednili.

V textu se vyskytuje celkem devět interpunkčních znamének (viz Tabulka 5).

Tabulka 5 Použitá interpunkční znaménka

Český název	**Interpunkční znaménko**	**Pinyin**	**Znaky**
Čárka	,	dòuhào	逗号
Tečka	。	jùhào	句号
Otazník	?	wènhào	问号
Vykřičník	!	tànhào	叹号

Středník	;	fēnhào	分号
Uvozovky	“ ”	yǐnhào	引号
Elipsa	……	shěnglüèhào	省略号
Dvojtečka	:	màohào	冒号
Čárka pro výčet	、	dùnhào	顿号

Zdroj: čínská terminologie převzata z (Biaodian fuhao, © 2017); vlastní zpracování

Pro účely segmentace jsme povídku převedli do abecedy pinyin dle normy GB/T 16159–2012 a po převedení text obsahuje 5 854 slov.

2.2.3 Han Han (Hán Hán 韩寒): *Život, jak mu rozumím* (Wǒ suǒ lǐjiě de shēnghuó 我所理解的生活)

Výběrový soubor *Život, jak mu rozumím* od čínského autora Han Han jsme zvolili z toho důvodu, že sloužil jako výběrový soubor již pro předešlý experiment, který zkoumal platnost MALu na čínských textech psaných v zjednodušených znacích (Motalová a Spáčilová 2013). V textu na něj budeme odkazovat jako Výběrový soubor 3. Článek byl publikován na blogu autora 20. června 2012 a obsahuje 2 641 znaků (bez mezer, s interpunkčními znaménky). Oproti předešlému výzkumu, kdy byl text segmentován dle grafického hlediska (Motalová a Spáčilová 2013, s. 98–100 a 103), jsme text převedli ze znaků do abecedy pinyin dle normy GB/T 16159–2012.

Jelikož autor pochází z pevninské Číny, originální text je psán v zjednodušených znacích. Článek je souvislý text, který sestává z 12 jasně vyznačených odstavců, a není přerušen žádnými grafy či obrázky s popisky. Většina čísel, které se v textu vyskytují, je zapsána pomocí čínských zjednodušených znaků. Pouze v jednom případě se objevilo arabské číslo 30. Toto číslo jsme stejně jako v případě výběrového souboru *Vítězství ženy* od Yu Hua ve výpočtech nezahrnuli. Z pohledu interpunkce je text napsán jednoduchým způsobem, operuje pouze se čtyřmi typy interpunkčních znamének, konkrétně čárka, tečka, vykřičník a dvojtečka (Tabulka 6). Přímá řeč dokonce není v textu vyznačena uvozovkami.

Tabulka 6 Použitá interpunkční znaménka

Český název	Interpunkční znaménko	Pinyin	Znaky
Čárka	,	dòuhào	逗号
Tečka	。	jùhào	句号
Otazník	?	wènhào	问号
Dvojtečka	:	màohào	冒号

Zdroj: čínská terminologie převzata z (Biaodian fuhao, © 2017); vlastní zpracování

Po převedení textu do abecedy pinyin dle normy GB/T 16159–2012 výběrový soubor obsahuje 1 624 slov.

2.2.4 Zhang Liqing (Zhāng Lìqīng): *Deníkové záznamy v pinyinu* (Pīnyīn Rìjì Duǎnwén)

Jako další výběrový soubor (Výběrový soubor 4) jsme zvolili kapitolu z deníku, jehož autorkou je Zhang Liqing (1936–2010). Autorka se narodila roku 1936 v pevninské Číně v provincii Shandong, ale v roce 1947 se její rodina přestěhovala na Taiwan. Studium absolvovala nejprve na Tainan Normal School a poté na National Taiwan University, kde získala tituly B.A. (1964) a M.A. (1966) v oblasti čínských studií. Ve studiu čínštiny pokračovala i na University of Washington, kde získala titul M.A. Zhang Liqing byla považována za výbornou učitelku čínštiny a vyučovala na několika prestižních univerzitách ve Spojených státech amerických. Byla také spoluzakladatelkou a spoluredaktorkou časopisu Xin Tang. Články publikované v tomto časopise jsou zvláštní svým zápisem, který je v podobě latinské transkripční abecedy Gwoyeu Romatzyh a pinyin. Časopis vycházel v letech 1982–1989 (Pinyin News: the blog of Pinyin.info 2010).

Deník *Pinyin Riji Duanwen* je netypický v tom, že čínská autorka pro zápis zvolila abecedu pinyin a v celém díle tak není použit žádný čínský znak. I když autorka nežila celý svůj život v pevninské Číně, díky stylu zápisu, který je ojedinělý, jsme její dílo zařadili mezi výběrové soubory. Deník vyšel v roce 2010. Zkoumaná kapitola nese název *Mihuo* a obsahuje 2 766 slov, což je 12 541 grafémů (bez mezer a s interpunkcí).

Jelikož byl soubor již původně zaznamenán v podobě latinky a slova jsou oddělena mezerami, nebylo tedy nutné přistoupit k manuální segmentaci slov. Text byl

ponechán v původní podobě. Autorčin styl zápisu v abecedě pinyin víceméně odpovídá normě GB/T 16159–2012 s výjimkou zápisů znaků 的, 得, 地 kdy znak 的 se přepisuje pouze jako d a znaky 得, 地 jako de. Oproti normě autorka v některých případech připojuje záložky k podstatným jménům (např. na ústech zuǐshang 嘴上; v ústech zuǐli 嘴里; ve městě chéngli 城里), dále připojuje záporku 不 bù k následnému slovesu nebo adjektivu (např. nechtít bùyào 不要; nebýt bùshì 不是; neodvážit se bùgǎn 不敢; nevydat bùchū 不出; nejasný buqīngchǔ 不清楚; nesnadný bùróngyì 不容易 ad.).

Struktura kapitoly je přehledná a každý odstavec začíná na novém řádku a je odskočený od okraje. Text není přerušen žádnými obrázky či grafy s popisky. Přímá řeč je v textu uvedena uvozovkami. Souvětí vždy začíná velkým písmenem a věty jsou dále členěny interpunkčními znaménky západního typu. Kupříkladu zde není použito čínské *kolečko* pro ukončení věty, ale tečka. V textu se vyskytuje celkem sedm interpunkčních znamének, viz Tabulka 7. V textu se objevilo jedno arabské číslo, ale stejně jako v případě ostatních výběrových souborů pro segmentaci nebylo zohledněno.

Tabulka 7 Použitá interpunkční znaménka

Český název	**Interpunkční znaménko**	**Pinyin**	**Znaky**
Čárka	,	dòuhào	逗号
Tečka	.	jùdiǎn	句点
Otazník	?	wènhào	问号
Vykřičník	!	tànhào	叹号
Středník	;	fēnhào	分号
Uvozovky	“ ”	yǐnhào	引号
Elipsa	…	shěnglüèhào	省略号

Zdroj: čínská terminologie převzata z (Biaodian fuhao, © 2017); vlastní zpracování

Jak jsme již uvedli výše naším cílem bylo zjistit, jestli zápis slov v pinyinu (tedy oddělování slov mezerami) odpovídá normě GB/T 16159–2012. Po srovnání všech výběrových souborů jsme zjistili, že až na několik výjimek i tento výběrový soubor

sleduje pravidla definovaná normou. Tyto výjimky budeme nadále analyzovat a zaměříme se na ně v kapitole 3, Interpretace získaných dat. Dále budeme pomocí kvantitativně-lingvistických metod ověřovat, jestli způsob segmentace, který je použit pro tento text, odpovídá ekonomizujícím zákonům a jestli by takto vytvořené ortografické slovo mohlo více odpovídat obecné definici čínského slova.

Výběrový soubor obsahuje 2 771 slov, což odpovídá 4 246 znakům.

2.2.5 Integrated Chinese Level 1 Part 2

Sestavu výběrových souborů doplňují učebnicové texty (Výběrový soubor 5), které jsou také originálně psané v pinyinu. Konkrétně jsme zvolili texty z učebnice *Integrated Chinese Level 1 Part 2*, lekce 19 a 20, což jsou poslední dvě lekce této učebnice a mělo by se tedy jednat o nejnáročnější úroveň. Učebnicové texty procházejí korekturou rodilých mluvčích, proto je považujeme za validní výběrové soubory, i když se transkripty v rámci různých učebnic mohou lišit. Celkový počet slov námi zvoleného učebnicového textu je 461, což je 1 902 grafémů (bez mezer). Jsme si vědomi, že oproti ostatním výběrovým souborům je délka podstatně kratší. Důvodem je, že čínské učebnice pracují s texty přepsanými do abecedy pinyin pouze do určité míry pokročilosti studentů a v dalších lekcích již s pinyinem nepracují a texty jsou sestaveny pouze ze znaků. Proto je většina textů v učebnicích pro začátečníky nebo mírně pokročilé kratší délky a není možné zvolit delší výběrový soubor. Na druhou stranu jsme se navzdory kratší délce rozhodli mezi výběrové soubory zařadit i tento text, protože věříme, že získáme dostatečné množství měřitelných hodnot.

Oproti ostatním výběrovým souborům je tento text napsán výrazně jednodušším stylem, což odpovídá právě povaze učebnicového textu. Výběrový soubor byl sestaven ze čtyř samostatných textů, které však na sebe obsahově navazují. Ve skutečnosti se jedná o jeden delší text, který je rozdělen na více částí, mezi kterými je vloženo vysvětlení gramatických jevů a slovníček. Tyto části nebyly do segmentace zahrnuty. Protože se jedná o jeden útvar, můžeme ho považovat za souvislý text.

Text jsme opět nesegmentovali, ale zachovali jsme jeho původní podobu. Oproti předešlému Výběrovému souboru 4, který byl také zaznamenán latinkou, se s normou shoduje mnohem méně. Na rozdíl od normy se v tomto textu vyskytují členitější slova, tzn. norma uvádí přepis do pinyinu jako jedno slovo, ale autoři učebnice zvolili oddělený zápis. Jedná se například o rozdělování slovesných přípon guo 过 a le 了, které jsou

psány odděleně od slovesa (šel qù guo 去过, uslyšel tīng le 听了, koupil mǎi le 买了, zarezervoval (si) dìng le 定了, odbavil tuōyùn le, našel zhǎo le 找了). Dále se jedná o slova, která by dle normy měla být zapsána dohromady, ale autoři je rozdělují (pracovat brigádně dǎ gōng 打工, bù déliǎo 不得了, vrátit se huí lai 回来, slevit dǎ zhé 打折, řídit kāi chē 开车, uvidět kàn jiàn 看见, nastoupit (do auta) shàng chē 上车).

Byla použita pouze základní interpunkční znaménka, konkrétně čárka, tečka a otazník (Tabulka 8).

Tabulka 8 Použitá interpunkční znaménka

Český název	**Interpunkční znaménko**	**Pinyin**	**Znaky**
Čárka	,	Dòuhào	逗号
Tečka	.	Jùdiǎn	句点
Otazník	?	Wènhào	问号

Zdroj: čínská terminologie převzata z (Biaodian fuhao, © 2017); vlastní zpracování

Výběrový soubor obsahuje 461 slov, což odpovídá 605 znakům. Zjistili jsme, že ve srovnání s ostatními výběrovými soubory autoři nereflektují v přepisu do pinyinu normu GB/T 16159–2012 a používají členitější slova. Stejně jako u předešlého výběrového souboru budeme zjišťovat, jestli tento způsob zápisu slov v pinyinu bude lépe odpovídat ekonomizujícím zákonům a jestli takto vytvořené ortografické slovo více odpovídá obecné definici čínského slova, viz kapitola 3.

Porovnání délek výběrových souborů

Ve srovnání s předešlými výzkumy jsme tentokrát výrazně navýšili nejenom počet zkoumaných souborů, ale také délky některých textů jsou výrazně delší. Kompletní přehled následuje v Tabulce 9.

Tabulka 9 Seznam výběrových souborů a jejich délky

Výběrový soubor	**Název výběrového souboru**	**Autor**	**Délka (ve znacích bez mezer)**	**Počet slov**
Výběrový soubor 1	*Kamarádi*	Yu Hua	7 527	4 571
Výběrový soubor 2	*Vítězství ženy*	Yu Hua	9 658	5 854
Výběrový soubor 3	*Život, jak mu rozumím*	Han Han	2 641	1 624
Výběrový soubor 4	*Deníkové záznamy v pinyinu*	Zhang Liqing	4 246[10]	2 771
Výběrový soubor 5	*Integrated Chinese Level 1 Part 2*	kolektiv autorů	605	461

2.3 Jazykové jednotky a jazykové úrovně

Precizní definování jazykových jednotek je důležitou součástí každého podobného experimentu. Aby byl výzkum prokazatelný, musíme přesně určit délku jednotlivých jednotek, tedy určit závislost jazykové jednotky na nižší jazykové hladině vůči jazykové jednotce na vyšší jazykové hladině. V našem případě klademe velký důraz na explicitní vymezení jazykových jednotek všech úrovní.

Jak uvádí Wimmer (2003), oproti přírodním vědám jsou jednotky ve společenských vědách definované vágněji a nejsou ani věčné, ani neměnné. Často se stává, že se definice určité jazykové jednotky může hodit pouze pro určitý typ jazyků. Proto platí, že žádná jazyková jednotka není správná nebo špatná, ale vždy záleží na cílech výzkumu (Wimmer et al. 2003, s. 18).

Jazykové jednotky mohou být navíc zvoleny podle různých kritérií; existuje např. kritérium fonetické, morfologické, syntaktické, grafické apod. Proto může pro jazykovou

[10] Výběrové soubory 4 a 5 byly zaznamenány formou latinky. Pokud bychom převedli slabiky na znaky, odpovídal by jejich počet číslu uvedenému v tabulce.

jednotku se stejným názvem existovat několik různých definicí (Wimmer et al 2003, s. 18). Obecně by jednotky měly být určeny podle těchto principů:

1) **Hranice**

   Definovaná jazyková jednotka musí být pomocí hranic dobře odlišitelná od okolí (Wimmer 2003, s. 18–19).

2) **Identita**

   Každá jednotka by měla mít svou identitu, tzn., že ji můžeme identifikovat nehledě na její synchronní variace nebo na trvalé diachronní změny. Identitu jazykové jednotce zajišťuje určitá formální nebo významová část (Wimmer 2003, s. 19).

3) **Integrace**

   Entita je jazykovou jednotkou pouze pokud působí na ostatní jednotky, anebo je ovlivňována jinými jednotkami (tj. když o ní můžeme vyslovit nějaké hypotézy) (Wimmer 2003, s. 19).

Již Marie Těšitelová (1987) upozorňovala, že při vymezování jazykových jednotek při jakékoli statistické analýze je důležité dodržovat tři zásady (Těšitelová 1987, s. 19):

1) jazykové jednotky musí být definovány jednoznačně
2) jejich pojetí by mělo být ve shodě s běžným pojetím v lingvistice
3) poté, co je přesně definujeme, důsledně dodržujeme jejich vymezení během celé práce

V případě našeho experimentu, který se zaměří na výzkum textů psaných v latince, jsme zvolili pět následujících jazykových jednotek:

grafém – slabika – slovo – klauze – souvětí

Vymezení těchto jednotek vychází ze struktury psaných textů a kombinuje psanou podobu čínského znakového písma a převod znakového písma do latinky. Nejnižší a dále nedělitelnou jednotkou psaného jazyka je *grafém*, který odpovídá jednomu písmenu latinské abecedy. Na další nejbližší vyšší jazykové úrovni se nachází *slabika.* Slabiku volíme z toho důvodu, protože čínský znak zpravidla odpovídá právě jedné slabice, což by mělo být zohledněno i po převodu znakového písma do latinky. Dále je nutné si uvědomit, že pinyin slouží jako nástroj převodu mluvené podoby čínštiny do podoby

grafické, neslouží jako nástroj pro zachycení přesné výslovnosti (Sehnal 1999, s. 85). Pinyin „neposkytuje jednoznačnou představu ani o výslovnosti slabik a slov (tu je v případě potřeby nutné uvádět v IPA), ani o jejich fonologické struktuře“ (Třísková 2010, s. 3). Z toho důvodu volíme slabiku jako další jednotku nadřazenou jazykové jednotce grafém. Jsme si vědomi, že slabika je fonetickou jednotkou, ale v našem případě ji zařadíme mezi zkoumané jazykové jednotky, protože pinyin pracuje právě s touto jednotkou a jeden znak přepisuje jako jednu slabiku. Nebudeme se tedy zabývat fonologickou stránkou slabiky, ale budeme na ni pohlížet jako odraz grafického hlediska znaků. Blíže o slabice viz kapitola 2.3.2.

Další vyšší jazykovou jednotkou je *slovo*, které v naší práci představuje hypotetickou jednotku, která je v textu vymezena dle pravidel normy GB/T 16159–2012 a přistupujeme k ní opět grafickým způsobem, tzn. hranice slov tvoří mezery určené touto normou (podrobněji rozebereme níže).

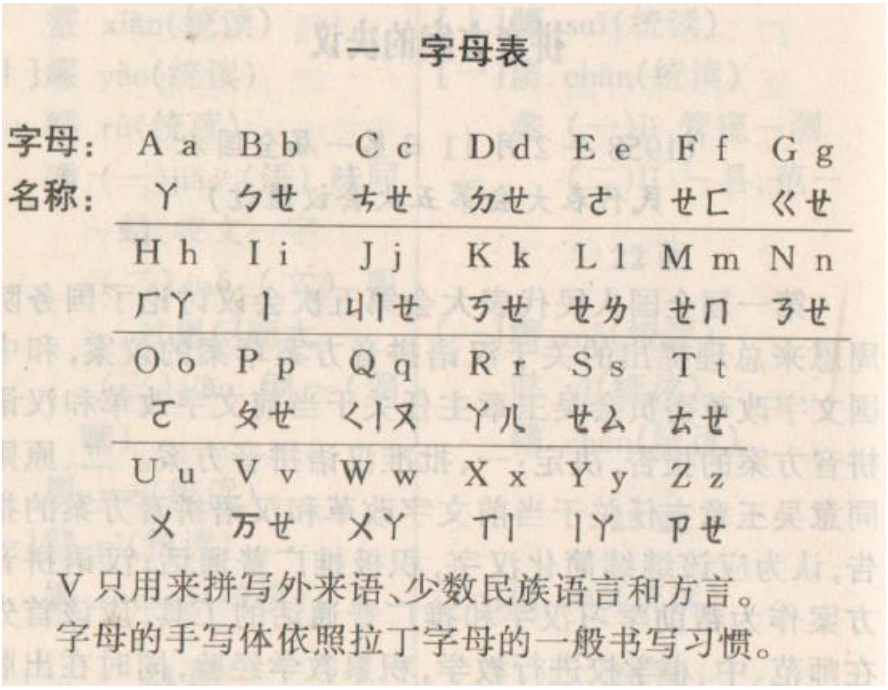
V hierarchii jazykových jednotek postupujeme přes *klauzi* až k nevyšší jazykové jednotce, kterou je *souvětí*. V první řadě jsme se snažili vymezit jazykové jednotky na základě grafického principu. V případě jazykové jednotek *klauze* jsme však byli nuceni toto pravidlo porušit, protože nebylo možné se opřít pouze o grafiku textu. Původním úmyslem bylo se z velké části opřít o interpunkci, jako ukazatele hranic klauzí, ale protože členění čínských textů je značně nesystematické, co se dělení interpunkčními znaménky týče, nebylo to možné. Proto jsme museli zvolit ještě další hledisko, a to syntaktické (viz kapitola 2.3.4). Nejvyšší jazyková jednotka *souvětí* je určena dle grafického principu na základě interpunkce.

Po vymezení jazykových jednotek je důležité, aby definice byly důsledně dodržovány na všech jazykových úrovních v průběhu celého experimentu.

2.3.1 Grafém (zìwèi 字位)

Grafém je obecně považován za nejmenší a dále již nedělitelnou jednotku psaného jazyka. Grafémy můžou mít různou podobu jako například písmena abeced, znaky nebo číslice. V tomto výzkumu mají grafémy podobu písmen latinské abecedy. Mezi grafémy se řadí i interpunkční znaménka, jelikož však slouží pouze k organizaci či k zdůraznění textu, nebyla pro účely segmentace zohledněna.

Hanyu pinyin fang'an (Hànyǔ pīnyīn fāng′àn 汉语拼音方; dále jen HPF), oficiální dokument z r. 1958, definuje pro přepis znaků do latinské abecedy 25 písmen, konkrétně to jsou písmena: a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v[11], w, x, y, z. Výslovnost těchto písmen vychází z fonetického systému čínštiny, a proto se liší od výslovnosti jiných jazyků používajících latinku. Například slovo 中国 (Čína) skládající se ze dvou čínských znaků 中 a 国 je do pinyinu přepsáno jako Zhongguo a skládá se tedy z osmi písmen (grafémů). Následující obrázek představuje tabulku z HPF, u které je pro výslovnost názvů písmen použit systém *Zhuyin zimu*. Pro bližší představu dále přikládáme tabulku, ve které uvádíme výslovnost písmen dle mezinárodní fonetické abecedy IPA.



Obrázek 2 HPF 1. Tabulka písmen (zìmǔ biǎo 字母表). Zdroj: HPF 1958, s. 296

Tabulka 10 Výslovnost písmen používaných v pinyinu dle mezinárodní fonetické abecedy IPA

písmeno	výslovnost IPA	písmeno	výslovnost IPA
Aa	[a]	Nn	[nɛ̲]
Bb	[pɛ̲]	Oo	[o]

[11] Písmeno ***v*** bylo v HPF začleněno pro účely zápisu výpůjček a jazyků národnostních menšin, nová norma GB/T 16159–2012 však již toto písmeno nepoužívá a pro zápis výpůjček a dalších slov využívá slabik, kterými se přepíše čínský znak zvolený pro přepis daného slova. Použití písmena ***v*** je možné při zadávání na klávesnici pro nahrazení písmena ü.

Cd	[ts‘ɛ̲]	Pp	[p‘ɛ̲]
Dd	[tɛ̲]	Qq	[tɕ‘iu]
Ee	[ɣ̲]	Rr	[ar]
Ff	[ɛ̲f]	Ss	[ɛ̲s]
Gg	[kɛ̲]	Tt	[t‘ɛ̲]
Hh	[xa]	Uu	[u]
Ii	[i]	Vv	[vɛ̲]
Jj	[tɕ̲iɛ̲]	Ww	[wa]
Kk	[k‘ɛ̲]	Xx	[ɕ̲i]
Ll	[ɛ̲l]	Yy	[ja]
Mm	[ɛ̲m]	Zz	[tsɛ̲]

Zdroj: Vlastní zpracování

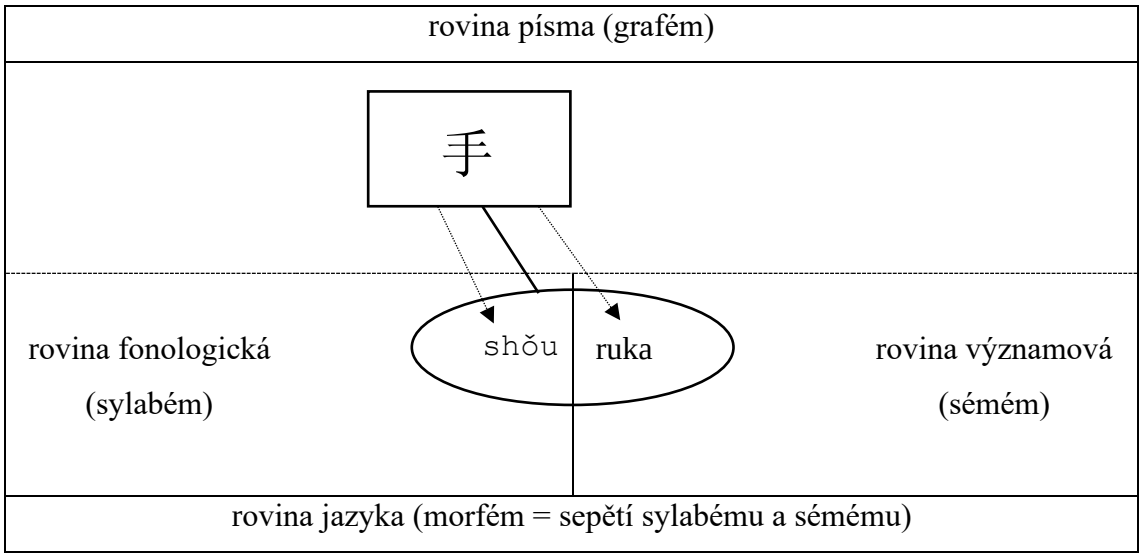
2.3.2 Slabika (yīnjié 音节)

Slabika má v čínském jazyce zásadní důležitost. Nejenom, že je výchozí jednotkou čínské výslovnosti, ale oproti evropským jazykům má vztah k významu a „…vystupuje jako hmotný nosič morfému. Je nejmenší jazykovou jednotkou, ve které dochází ke spojení významu s určitou výslovností (yì yīn hétí 意音合体)“ (Třísková 2010, s. 47). V čínštině se tedy může realizovat slabika pouze v tom případě, že ji můžeme asociovat s nějakým významem (Sehnal 2002, 14). Počet slabik v systému současné čínštiny je omezený a nelze tak vytvářet nové libovolné slabiky. Pokud například čínština přejímá slova cizího původu do své slovní zásoby, dochází ke zkomolení tak, aby se příslušné slovo dalo rozčlenit na čínské slabiky, které se dají zapsat nějakými již existující čínskými znaky bez ohledu na jejich význam.

Obecně můžeme říct, že až na naprosté výjimky odpovídá čínský znak (grafém[12]) jednomu morfému, který se skládá z nejmenší významové jednotky (sémém) a tato jednotka je vázána na určitý zvuk (sylabém), viz Obrázek 3. Čínština je jazyk slabičný a jeho základní fonologickou jednotkou je slabika, která má pevně danou stavbu (sylabém) a většina morfémů odpovídá na zvukové úrovni jedné slabice (Zádrapa 2009, s. 37). Znak tedy není vázán pouze na zvuk nebo význam, ale vztahuje se k morfému jako celku.

[12] V této souvislosti grafém neoznačuje písmeno latinské abecedy, ale čínský znak

Čínské písmo tak můžeme označit za morfemografické. Tento vztah ilustrujme například na znaku shǒu 手 s významem *ruka*:

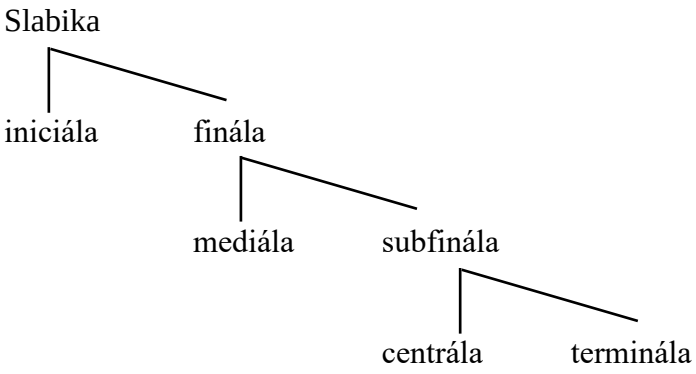


Obrázek 3 Charakteristický vztah mezi čínským znakem, tj. grafémem, a jazykovou jednotkou morfémem. Zdroj: Zádrapa 2009, s. 37

V praxi to vypadá tak, že jeden znak vyslovíme jako jednu slabiku. Současná standardní čínština operuje přibližně celkem s 400 slabikami, pokud k nim přičteme i slabiky rozlišené tóny, dostaneme něco málo přes 1 200 slabik (Třísková 2010, s. 342; Zádrapa 2009, s. 43).

Výjimkami z pravidla jedna slabika = jeden znak může být např. erizace, která je typická pro severní dialekty. Jedná se o připojování koncového -r na konec čínské slabiky: wánr 玩儿, nàr 那儿, zhèr 这儿, xiǎoháir 小孩儿. Jak je vidět z příkladů, slovo zapsané dvěma znaky se vysloví jednou slabikou a slovo zapsané třemi znaky se vysloví dvěma slabikami atd. Zápisu finály er 儿 se věnuje i dokument HPF, který přímo uvádí, že pokud se jedná o samostatnou slabiku er (např. *ucho* ěr 耳) zapíše se slabika jako „er“. Pokud se jedná o erizaci, připojuje se pouze koncové 儿 „r“ za slabiku. Erizace je typická pro severní dialekty, ale kromě toho se podílí se i na slovotvorbě, může plnit významotvornou funkci nebo naznačovat familiárnost (Kane 2009, s. 152–153). Je však třeba poznamenat, že v psaném projevu se přípona často vypouští a erizovaná výslovnost může být realizovaná až při čtení z kontextu (Zádrapa 2009, s. 41–42). Pro účely našeho experimentu jsme připočítávali koncové -r k předešlé slabice a dohromady tak dva znaky byly započítány jako jedna slabika.

Pro náš výzkum je zásadní zejména zápis slabiky jakožto sekvence hlásek. Uveďme alespoň ve stručnosti stavbu čínské slabiky, se kterou pracuje abeceda pinyin. HPF výslovně nezmiňuje popis struktury slabiky, představuje pouze její komponenty, tj. iniciálu, finálu a tón, ale z tabulek uvedených v tomto dokumentu lze vyvodit, že pinyin přijímá tradiční pohled na slabiku, tedy rozdělení slabiky na iniciálu a finálu, která se dále dělí dle následujícího schématu (Třísková 2010, s. 57, 113):



Obrázek 4 Tradiční model čínské slabiky. Zdroj: Třísková 2010, s. 113

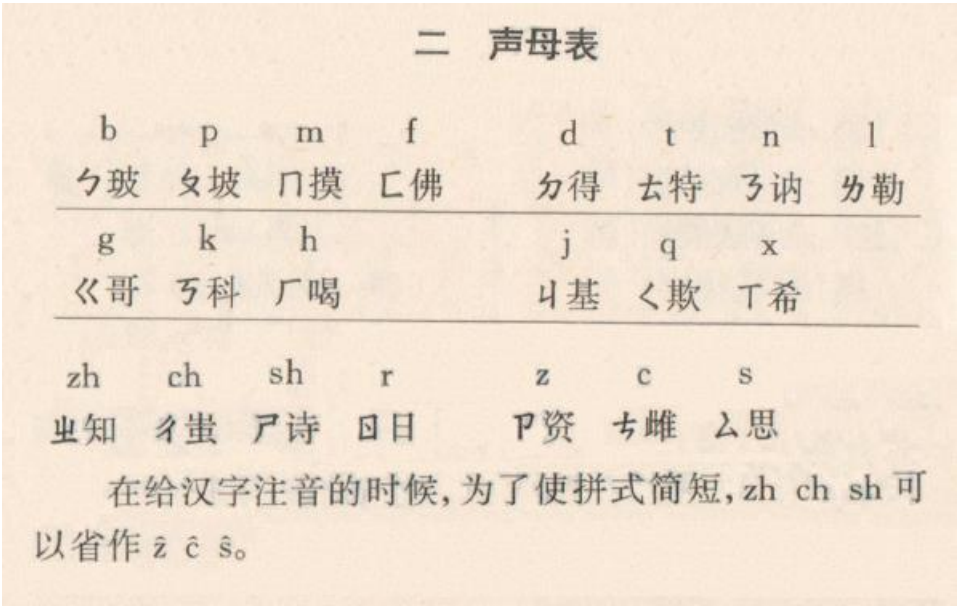
Kombinace hlásek v rámci slabiky jsou v čínském jazyce značně omezené. „Souhlásky se mohou objevit pouze na začátku slabiky jako tzn. iniciály. Zbytek slabiky, tzn. finála, je tvořen jednoduchou samohláskou (a, e, i, u, ü), nebo dvojhláskou (ai, ei, ao, ou). Samohláskám a, e, jakož i vyjmenovaným dvojhláskám může ještě předcházet tzv. mediála, tj. kratičké i, u nebo ü. Jednoduché samohlásky a, e (ať již s mediálou, nebo bez ní) mohou být navíc zakončeny nosovkou: přední nosovkou -n, nebo zadní nosovkou -ng“ (Švarný 1999, s. 42). Z pohledu pinyinu je jedinou povinnou složkou centrála, kterou může být pouze samohláska (Třísková 2010, s. 114). Na jednotlivých pozicích se mohou objevit následující inventáře:

iniciála (shēngmǔ 声母)	finála (yùnmǔ 韵母)		
	mediála (yùntóu 韵头)	centrála (yùnfù 韵腹)	terminála (yùnwei 韵尾)
b, p, m, f *d, t, n, l* *z, c, s* *zh, ch, sh, r*	*i, u, ü*	*a, o, e, i, u, ü,* (ê), (*er*)	*i, u* *n, ng*

j, q, x *g, k, h*			

Obrázek 5 Inventáře komponentů iniciála, mediála, centrála a terminála přijímané pinyinem. Zdroj: Třísková 2010, s. 114.

Dokument HPF uvádí celkem pět tabulek, z nichž první vyjmenovává písmena, které je v pinyinu možné používat (viz kapitola 2.3.1 Grafém). Následující dvě tabulky uvádějí iniciály a finály, čtvrtá tabulka obsahuje čtyři diakritické značky pro označení tónů a poslední pátá tabulka poukazuje na použití oddělovacího znaménka, tedy apostrofu, který se užívá pro oddělení slabik, kdy mezi nimi není jasná hranice. Oddělovací znaménko také napomáhá k lepší orientaci při zápisu víceslabičných slov. Jelikož v naší segmentaci tón necháváme stranou, níže krátce představíme pouze tabulky iniciál a finál a oddělovací znaménko.



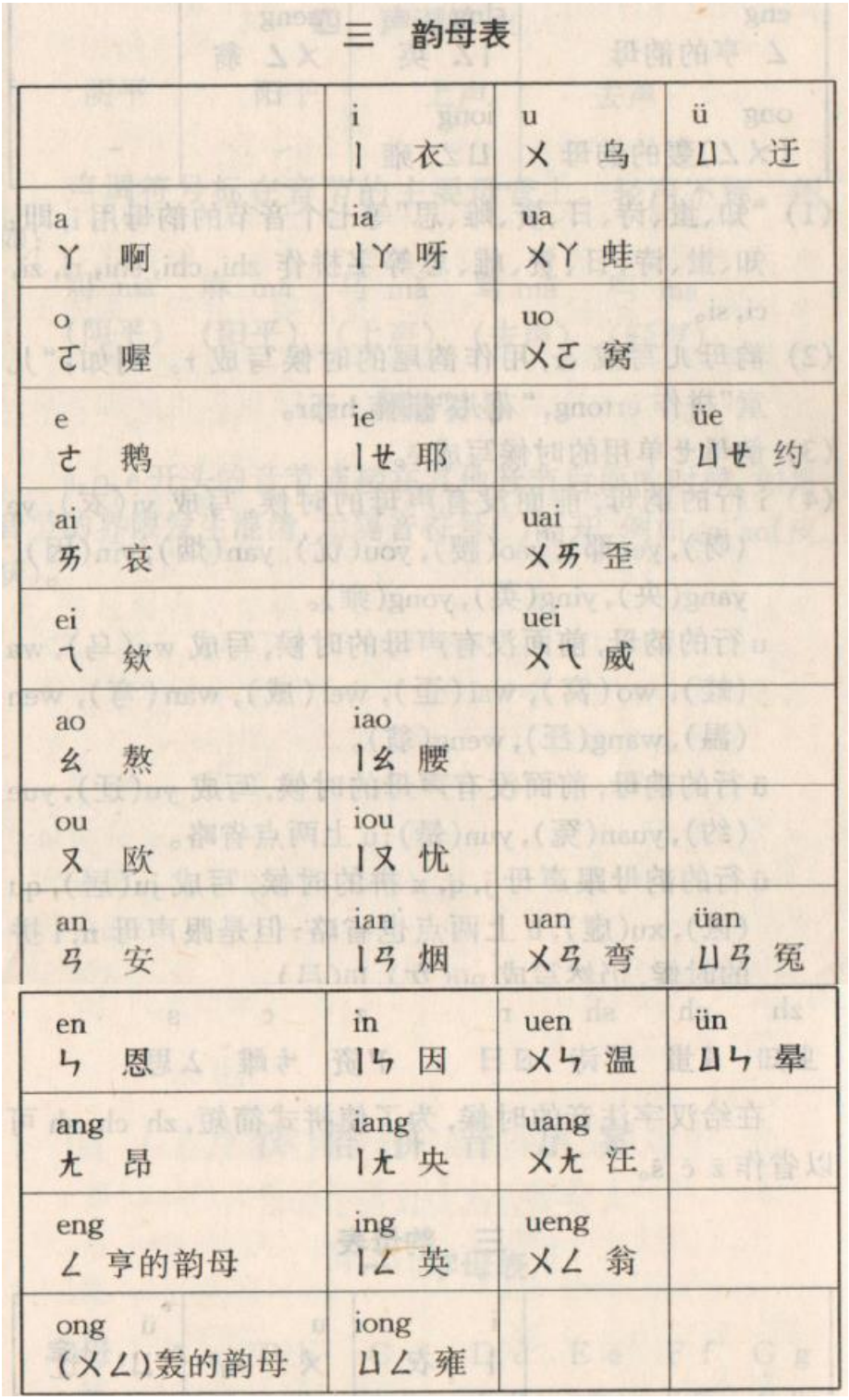
二　声母表

b	p	m	f	d	t	n	l
ㄅ玻	ㄆ坡	ㄇ摸	ㄈ佛	ㄉ得	ㄊ特	ㄋ讷	ㄌ勒
g	k	h		j	q	x	
ㄍ哥	ㄎ科	ㄏ喝		ㄐ基	ㄑ欺	ㄒ希	
zh	ch	sh	r	z	c	s	
ㄓ知	ㄔ蚩	ㄕ诗	ㄖ日	ㄗ资	ㄘ雌	ㄙ思	

在给汉字注音的时候，为了使拼式简短，zh ch sh 可以省作 ẑ ĉ ŝ。

Obrázek 6 HPF Tabulka iniciál (`shēngmǔ biǎo` 声母表). Zdroj: HPF 1958, s. 296–297

Tabulka iniciál obsahuje 21 položek a stejně jako u první tabulky z dokumentu HPF, která představuje latinská písmena (viz Obrázek 2), je i u těchto položek pro výslovnost zvolen zápis v systému *Zhuyin zimu*. Navíc jsou zde uvedeny i znaky, jejichž čtení začíná danou iniciálou.



三　韵母表

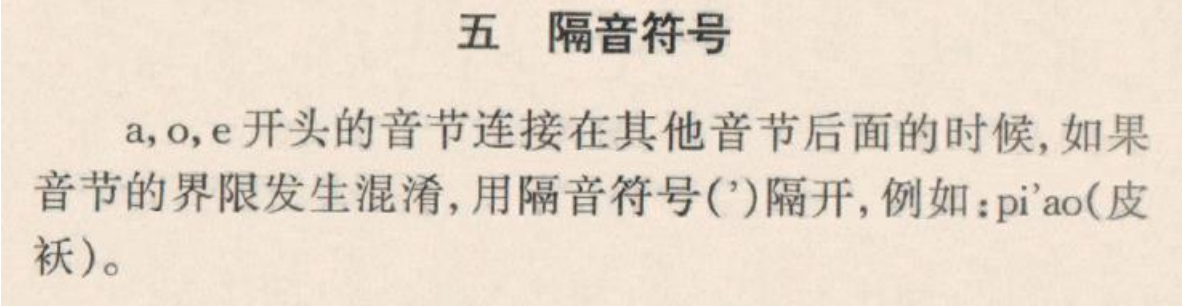
	i ㄧ 衣	u ㄨ 乌	ü ㄩ 迂
a ㄚ 啊	ia ㄧㄚ 呀	ua ㄨㄚ 蛙	
o ㄛ 喔		uo ㄨㄛ 窝	
e ㄜ 鹅	ie ㄧㄝ 耶		üe ㄩㄝ 约
ai ㄞ 哀		uai ㄨㄞ 歪	
ei ㄟ 欸		uei ㄨㄟ 威	
ao ㄠ 熬	iao ㄧㄠ 腰		
ou ㄡ 欧	iou ㄧㄡ 忧		
an ㄢ 安	ian ㄧㄢ 烟	uan ㄨㄢ 弯	üan ㄩㄢ 冤
en ㄣ 恩	in ㄧㄣ 因	uen ㄨㄣ 温	ün ㄩㄣ 晕
ang ㄤ 昂	iang ㄧㄤ 央	uang ㄨㄤ 汪	
eng ㄥ 亨的韵母	ing ㄧㄥ 英	ueng ㄨㄥ 翁	
ong (ㄨㄥ)轰的韵母	iong ㄩㄥ 雍		

Obrázek 7 HPF Tabulka finál (`yùnmǔ biǎo` 韵母表). Zdroj: HPF 1958, s. 297–298

Tabulka finál má 35 položek, pro výslovnost je také zvolen systém *Zhuyin zimu*. Tabulka dále ukazuje příklady znaků, jejichž čtení obsahuje danou finálu.

Přísná pravidla ve fonologické stavbě slabiky zaručují jednoznačné určení slabičné hranice. Případy, kdy by mohlo dojít k nejednoznačnému zápisu, řeší HPF

potažmo pinyin jednak apostrofem (město 西安 se zapíše jako `Xi'an`, nikoli jako `Xian`), jednak využitím písmen ***y, w*** na začátku slabik s neobsazenou iniciálou (číslo 5 se v pinyinu zapíše jako `wu`, nikoli jako `u`; Třísková 2010, s. 97–98).



五　隔音符号

a, o, e 开头的音节连接在其他音节后面的时候，如果音节的界限发生混淆，用隔音符号(')隔开，例如：pi'ao(皮袄)。

Obrázek 8 Oddělovací znaménko (`géyīn fúhào` 隔音符号). Zdroj: HPF 1985, s. 299

Slabika v pinyinu je tudíž precizně definovaná a existuje přesný počet slabik, které se v moderním čínském jazyce používají. Nejkratší slabiky mohou být tvořeny jedním grafémem (např. citoslovce `a` 啊) a nejdelší slabiky šesti grafémy (např. slovo *postel* `chuáng` 床). Jak již bylo zmíněno v kapitole 2.1.2.2, při segmentaci nebudeme zohledňovat tóny slabik, jelikož neovlivňují délku slabiky, kterou měříme v grafémech.

2.3.3 Slovo (`cí` 词)

V některých kvantitativně-lingvisticky zaměřených pracích se setkáváme s jazykovou jednotkou slovo, aniž by bylo jasně definováno, co přesně touto jednotkou autor myslí a jakým způsobem ji vymezuje. Např. Hou et al. (2019, s. 4) pro svůj výzkum vágně definují slovo jako: „Segmenty, které jsou v textu ohraničené mezerami a segmentované čínským systémem lexikální analýzy, považujeme jako operačně definovaná slova.“[13]. Chen a Liu (2014, s. 83) vymezují slovo následujícím způsobem: „Protože v čínštině nejsou mezi slovy žádné mezery, použili jsme k segmentaci textů nejoblíbenější nástroj pro segmentaci čínských slov ICTCLAS 2008…“[14] ICTCLAS pro segmentaci slov využili ve svém výzkumu také Renkui Hou a jeho kolegové (2017). Studie se často odkazují na použití různých korpusů, ale jak poukazuje Qiao s kolegy

[13] *We take the segments delineated by blank spaces in the texts, segmented by the Chinese lexical analysis system, as operationally defined words.* Vlastní překlad autorky.

[14] *Since there are no spaces between words in Chinese, we used the most popular Chinese word segmentation tool ICTCLAS 2008 to segment the texts, ...*Vlastní překlad autorky.

(2010) ideální korpus ať už vytvořený manuálně, nebo automaticky, neexistuje a „absence praktického standardu pro definování čínských slov vede k problémům s nekonzistencí při vytváření korpusu[15]“ (Qiao et al. 2010, s. 143, 145). Na slovo však můžeme pohlížet z různých perspektiv a nejednoznačné definování jazykové jednotky může přinést nerelevantní výsledky. Vždy záleží, pro jaký účel má jednotka sloužit, zda zkoumáme fonetické, gramatické, sémantické a historické (etymologických) aspekty (Wimmer 2003, s. 87). Například Jerome Packard (2000, s. 8–20) rozlišuje po zohlednění různých hledisek osm druhů slov: slovo ortografické, sociologické, lexikální, sémantické, fonologické, morfologické, syntaktické a psycholingvistické.

Jak již bylo uvedeno výše, pro kvantitativně-lingvistický výzkum není žádná definice špatná, je třeba vždy zohlednit jaká jednotka je nejvhodnější pro konkrétní výzkum a zvolenou definici pečlivě dodržovat.

Jedním z faktorů při definování slova může být, zda jednotka pochází z jazyků flektivních či analytických. Ve flektivních jazycích může být na slovo pohlíženo jako na jednotku jazykového systému (lexém), což je základní podoba slova, kterou nalezneme ve slovníku a reprezentuje celou množinu svých slov (lexém *kniha* může mít v textu podobu *knihou*, *knihy*, *knihu* apod., lexém *číst* může mít podobu *četl jsem*, *četl bych* apod.; Těšitelová 1987, s. 12). Příkladem lexikálního slova může být *Čínská lidová republika* (Zhōnghuá rénmín gònghéguó 中华人民共和国), které z lexikálního pohledu představuje jedno slovo, avšak skládá se ze tří fonologických slov (Sproat et al. 1996, s. 379). Další možností je slovo vnímat jako slovoformu, tedy morfematickou a morfologickou jednotku skládající se ze série morfů (Benešová 2011, s. 35).

Slovo je možné vymezit také graficky jako „skupinu písmen mezi dvěma mezerami“, tzn. ortografické slovo. Toto pravidlo se snadněji uplatní u jazyků analytických (izolačních), mezi které se řadí i čínština. V čínském jazyce jsou vztahy mezi větnými členy vyjádřeny pomocí slovosledu a tzv. funkčních slov. Oproti flektivním jazykům čínština nepoužívá skloňování nebo časování, forma slov zůstává stále stejná. V podstatě by se dalo říci, že čínština je pro grafický způsob definice ideální, protože tvary slov se ve větách nemění a používají se základní podoby slov, které najdeme ve slovníku[16].

[15] The lack of a practical standard for Chinese words leads to inconsistency problems in corpus construction. Vlastní překlad autorky.

[16] Výjimkou jsou zde jsou např. předložky, záložky či slovesné přípony, které se k základnímu tvaru slova přidávají až v textu.

Avšak v případě čínštiny zde narážíme na překážku. Čínský text je jednolitý a čínské znaky, kterými je text psán, stojí vedle sebe a není mezi nimi žádná mezera. Slovo tak není typickou jednotkou čínského písma, protože jeho hranice nejsou žádným způsobem vyznačeny (Norman 1988, s. 155). Definici založenou na grafice není možné aplikovat tak jednoduše, jak by se zdálo. Není to ovšem neřešitelný problém, protože text psaný ve znacích je možné převést do latinky a následně použít grafický princip pro rozlišení slov. Převod je možné provést různými způsoby, které nastíníme níže. Nejprve se však podívejme na problematiku slova v moderní čínštině.

2.3.3.1 Koncept slova v moderní čínštině

Dle Packarda (2000, s. 14–15), slovo není v čínštině jasnou a intuitivní jednotkou. Pojem, který lze v čínském jazyce a kultuře považovat za jasný, je zì 字, který se všeobecně používá ve dvou významech: v mluveném jazyce označuje morfém a v psaném jazyce čínský znak. Avšak většina čínských mluvčích mezi těmito významy nerozlišuje a pojem zì 字 pro ně znamená jednu a tutéž věc. Jerome Packard (2000) stejně jako Huang a Xue (2012) tento pojem označují jako sociologické slovo. Yuan Ren Chao popisuje sociologické slovo jako „… typ jednotky, která velikostně stojí na pomezí mezi fonémem a větou a široká nelingvistická veřejnost si ji uvědomuje, mluví o ní, má pro ni každodenní termín a prakticky se jí různým způsobem zabývá[17]“ (Yuan Ren Chao 2011, s. 159).

Čínština v současné době používá nejenom jednoznaková slova, ale i slova skládající se ze dvou a více znaků (morfémů). Pro pojem *slovo* existuje v čínštině výraz cí 词, které může označovat jeden znak (morfém), ale i spojení více znaků (morfémů), které tvoří jeden celek. Z lingvistického hlediska bychom dle Packarda tento výraz mohli označit jako syntaktické slovo (2000, s. 15). Jedno slovo cí 词 se může skládat ze dvou i více znaků (morfémů) zì 字. Např. slovo *auto* (qìchē 汽车) se skládá ze dvou znaků (morfémů) *pára* (qì 汽) a *vůz* (chē 车). Jak je vidět z příkladu, tyto dva znaky tvoří jeden celek, tj. slovo, a není možné žádnou jeho část nahradit jiným znakem, aniž bychom změnili význam slova. Přestože výraz cí 词 bývá pro termín *slovo* používán, bohužel

[17] ... type of unit, intermediate in size between a phoneme and a sentence, which the general, nonlinguistic public is conscious of, talks about, has an everyday term for, and is practically concerned with in various ways. Vlastní překlad autorky.

neexistuje přesná definice, která by určovala hranici čínského slova (Packard 2000, s. 15–16).

Kromě pojmu slovo (cí 词) je v čínštině nutné pracovat také s pojmem slovní spojení (cízǔ 词组). Obecné definice *slovního spojení* v čínštině mohou znít následovně: „Slovní spojení je jazyková jednotka složená ze dvou nebo více slov, utvořená podle určitých gramatických pravidel a vyjadřující určitý význam“[18] (GB/T 13715–92, s. 1). Další definice popisuje slovní spojení jako „ustálené spojení dvou a více slov, které se opakovaně používá v různých větách. Slovo i slovní spojení může vytvářet jednoslovné či krátké věty, …, ale nelze říci, že slovo či slovní spojení je věta. Slovní spojení je v čínštině z hlediska gramatické hierarchie považováno za prostředníka mezi slovy a větami[19]“ (Cizu, © 2022).

Ne vždy je jednoznačné, jestli lze jednotku považovat za slovo či za slovní spojení. Například Fu Huaiqing (2020, s. 1, 3) navrhuje definovat čínské slovo pomocí tří charakteristik: slovo má daný význam, má danou syntaktickou charakteristiku (je nejmenší jednotkou, která může být vyslovena samostatně, nebo nejmenší jednotkou, ze které se skládá věta) a má obecně ustálenou fonologickou podobu. Fu dále uvádí metody, kterými je možné určit, zda se jedná o slovo či o slovní spojení (2020, s. 4–8):

1. Pokud se jednotka může říct samostatně a může sama tvořit odpověď na otázku, je to slovo (Fu 2020, s. 4). Například: *Co je to?* (Zhè shì shénme? 这是什么？) ***Hruška.*** (**Lí. 梨。**). Slovo *hruška* je možné považovat za slovo. Tato metoda však není dostačující pro všechny případy. Na otázku lze totiž odpovědět nejen slovem, ale také slovním spojením. Např. *Co jsi koupil?* (Nǐ mǎi shénme? 你买什么？) *Papír. / Bílý papír.* (Zhǐ. / **Bái zhǐ.** 纸 / **白纸**). I když výraz *bílý papír* (bái zhǐ 白纸) můžeme použít v odpovědi na otázku, jedná se o slovní spojení a ne slovo (blíže bod 3). Dalšími příklady, kdy se jedná o slovní spojení, které lze rozdělit na jednotlivá slova, jsou např. *ne dobrý* (bù hǎo 不好), *ne-koupil* (bù mǎi 不买) apod.

2. Pokud se jednotka nedá říct samostatně, ale může fungovat jako větný člen, je to slovo (Fu 2020, s. 5). Jedná se o jednoslabičná slova jako například *mužský* (nán 男),

[18] 有两个或两个以上的词，按一定的语法规则组成，表达一定意义的语言单位。Vlastní překlad autorky.

[19] 词组是指两个或多个词构成一定的组合关系，又经常在不同的句子里一起使用着的固定语句片段；词、词组都可以单独成句子，这句子是独词句或短句，但不能说词、词组是短句；在汉语语法中，从语法层级上看，词组介于词和句的中间。Vlastní překlad autorky.

zlatý (jīn 金) apod. Na otázku *Její dítě je chlapec nebo dívka?* (Tā de háizi shì nán de háishi nǚ de? 她的孩子是男的还是女的？) Při odpovědi nelze použít pouze samostatně nán 男, ale je třeba říci ve spojení s přivlastňovacím de 的, tedy nán de 男的。I přesto lze znaky jako nán 男, jīn 金 apod. považovat za slova, protože mohou být použita v různých slovních spojeních a vstupují do věty jako větný člen. Např. *pánské boty* (nán píxié 男皮鞋), *zlatý prsten* (jīn jièzhi 金戒指) apod. Kromě toho do této kategorie Fu Huaiqing (2020, s. 5) řadí i jména míst, např. nèi 内, wài 外, lǐ 里 ad. (*ve třídě* jiàoshì nèi 教室内, *mimo školu* xuéxiào wài 学校外…), která také nelze použít samostatně, ale jsou slovy.

3. Pokud se jazyková jednotka skládá z více složek a mezi jednotlivé složky lze vložit jiný komponent, aniž by došlo ke změně významu, jedná se o slovní spojení (Fu 2020, s. 6). Fu (2020, s. 6–7) na níže uvedených příkladech ilustruje rozpad slovního spojení na jednotlivá slova:

bílý papír bái zhǐ 白纸→ *bílý papír* bái de zhǐ 白的纸

dosyta se najíst chī bǎo 吃饱→ *dosyta se najíst* chī de bǎo 吃得饱

Dále Fu (2020) uvádí případy, kdy mezi jednotlivé složky nelze vložit komponent, aniž by se změnil původní význam, a celek tedy tvoří slovo (nikoliv slovní spojení):

všichni dàjiā 大家→ nelze vytvořit tvar *dà de jiā *大的家, význam by se změnil na *velká rodina*

mořské řasy hǎidài 海带→ nelze vytvořit tvar *hǎi de dài *海的带, význam by se změnil na *pásmo moře*

rozlehlý kuòdà 扩大→ nelze vytvořit tvar *kuà de dà *扩得大, význam by se změnil na *rozšířeno na velkou plochu*

Nutno podotknout, že tuto metodu také není možné aplikovat na všechny případy. Existují slovesa, mezi které nelze vložit další komponent, a přesto jsou slovními spojeními a nikoliv slovy, např. *Honem pojď*. kuài zǒu 快走, *jen co přišel* gāng lái 刚来 apod.

4. Pokud se zbavíme všech jednotek ve větě, které lze vyslovit samostatně a lze je použít jako hlavní složku věty, zbývající jednotky, které nelze vyslovit samostatně a nejsou součástí jiného slova, jsou slova (Fu 2020, s. 8). Jedná se o funkční slova jako například spojky, partikule apod.

Výše uvedená pravidla naznačují možné způsoby, jak lze přistupovat k určení slova a slovního spojení v čínském textu. Pro naši práci tedy nehrají zásadní roli a při segmentaci se jimi nebudeme řídit, protože se ve všech aspektech neshodují s pravidly normy GB/T 16159–2012. Například se liší v možnosti připojování přivlastňovacího slovce de 的 k osobnímu zájmenu, což dle Fua není možné, ale norma tento způsob v určitých případech umožňuje.

2.3.3.2 Způsoby segmentace čínského textu na slova

Texty psané čínskými znaky je možné převádět i automaticky za využití různých softwarů, které čínský text automaticky rozsegmentují na slova, např. výše zmíněný ICTCLAS 2008, který se opírá o rozsáhlý lexikon a Markovův model a začleňuje tokenizaci slov, identifikaci pojmenovaných entit, rozpoznávání neznámých slov a označování částí řeči (Corpus annotation 2022).

Při automatické segmentaci slov psaného čínského textu však existují různá úskalí, se kterými je třeba se vypořádat. První z nich je nejednoznačnost, např. spojení 个人 lze chápat jednak jako slovo složené ze dvou znaků gèrén *individuální (osoba)* nebo jako dvě samostatná jednoznaková slova ge rén *jeden člověk* (Huang a Xue 2012, s. 495–496). Kromě toho může být obtížné rozpoznat neznámá slova v textu (např. vznik nových slov odvozováním či skládáním, popřípadě neologismy a transliterace cizích slov; Huang a Xue 2012, s. 495–496). Automatická segmentace se musí dále vypořádat i s několika případy homografů, které se v čínštině vyskytují. Tzn. jeden znak může být vysloven více způsoby a mít různé významy. Např. znak 地 lze vyslovit jako de (funkční slovo) nebo jako dì ve významu *země*, *půda*.

Huang a Xue (2012) popisují vývoj přístupů automatické segmentace čínského slova. První a jednou z nejpoužívanějších a nejintuitivnějších metod byla Metoda porovnávání vzorů (Pattern Matching Approach: Dictionary Lookup). U této metody by měl být počítač schopen rozčlenit čínskou větu, která se skládá z řetězců znaků, na slova pomocí vyhledávání slov ve slovníku. Jak jsme již naznačili výše, prvním úskalím této metody je nejednoznačnost při rozlišování hranic slov. V určitých případech se může jednat o řetězec skládající se z jednoho slova (kěyǐ 可以), avšak v jiných případech řetězec složený ze dvou slov (kě yǐ 可以). Pro řešení nejednoznačnosti lze použít metodu nazývanou *algoritmus maximální shody*, kdy se počítač snaží najít nejdelší řetězec čínských znaků, který se shoduje se slovem ve slovníku. Tím, že je řetězec dvou

znaků obsažený ve slovníku, je pravděpodobnější, že se bude jednat o slovo, a ne o dvě samostatná slova. Jelikož však algoritmus nezohledňuje kontext, vždy upřednostní dvouznakové slovo před dvěma jednoznakovými (Huang a Xue 2012, s. 496). Další omezení metody porovnávání vzorů tkví v nemožnosti vytvořit slovník, který by obsahoval veškerá možná čínská slova (McCallum a Feng 2003, s. 1).

Druhá metoda aplikující stochastický přístup se nazývá Statistická metoda a identifikuje slovo pomocí pevnosti vnitřní vazby mezi dvěma znaky (Statistical Approach: Strength of Internal Binding). Jedním z prvních statisticky zaměřených přístupů používá k měření síly vazby v řetězci znaků *Mutual Information (MI, vzájemnou informaci)*. Čistě statistické metody neporovnávají výsledky se slovníky, ale pracují v rámci korpusu. Páry znaků, které se opakují, mají vysokou hodnotu MI, pokud je vysoká pravděpodobnost, že se tyto dva znaky vyskytnou společně, a zároveň pokud je nízká pravděpodobnost, že se vyskytnou samostatně. Nejvyšší hodnota MI je, pokud se dva znaky vyskytnou společně ve všech případech. Nevýhodou čistě statistického přístupu je, že dokáže poměrně přesně vymezit dvouznaková slova, ale má problém určit slova víceznaková (Huang a Xue 2012, s. 497–498).

Třetí metoda nazývaná Označování znaků (Character Tagging) zahrnuje označování znaků podle pozice, ve které se v rámci slova vyskytují (Huang a Xue 2012, s. 498–500). Každému znaku z korpusu se nejprve ručně přiřadí jedna či více následujících pozic: LL (vlevo: znak začíná slovo, ale sám o sobě slovem není), RR (vpravo: znak se nachází na pravé straně slova, ale sám o sobě slovem není), MM (uprostřed: znak je uprostřed slova) a LR (slovo samo o sobě; Xue 2003, s. 35). Např. marker množného čísla pro osoby men 们 může mít značení RR. K řešení nejednoznačnosti při označování poloh je vhodné tento model kombinovat se Strojovým učením, které efektivně využívá kontextové informace. Výhodou metody Označování znaků je, že se umí vypořádat i se slovy, která nejsou obsažena ve slovníku. Počet znaků v čínštině zůstává poměrně konstantní a většinou nepředpokládáme, že se objeví nový znak. Pro neologismy pak s nejvyšší pravděpodobností bude použit již existující znak, který bude stát na obvyklém místě (Huang a Xue 2012, s. 498–500).

Poslední a nejnovější představovanou metodou je Rozhodování o hranicích slov (Word Boundary Decision). Oproti předchozím metodám se při segmentaci nemusí klasifikovat slova nebo znaky, ale klasifikují se pouze intervaly mezi dvěma znaky. V segmentovaném textu jsou všechny intervaly mezi znaky označeny jako hranice slova nebo jako hranice znaku. Metoda tedy navrhuje omezit úlohu na binární rozhodování

o jedné jednotce, což usnadňuje a zrychluje segmentaci. Každá hranice (interval mezi znaky) je obklopena řetězcem znaků, které slouží pouze k určení kontextu a není důležité, jestli řetězec znaků tvoří slovo nebo ne. V prvním kroku této metody se odvodí sady *n*-gramů a vypočítá se pravděpodobnost, jestli je interval hranicí slova vzhledem k danému kontextu (Li a Huang 2009, s. 727). V druhém kroku jsou pravděpodobnosti použity pro generování vektorů, které jsou označeny buď jako Typ 0 (hranice znaku), nebo jako Typ 1 (hranice slova) v závislosti na klasifikaci intervalu v korpusu (Huang et al. 2008, s. 4). V případě, že nelze prvku přiřadit odpovídající hodnotu z korpusu, použije se náhodná pravděpodobnost 0,5.

Z výše uvedeného vývoje přístupů k automatické segmentaci je patrné, že nové modely snižují složitost segmentačního procesu (Huang a Xue 2012, s. 500). Všechny zmíněné metody segmentace čínského textu pracují s texty psanými čínskými znaky a při segmentaci by měly zohledňovat národní normu GB/T 13715–92 *Contemporary Chinese language word segmentation specification for information processing*, která specifikuje pravidla pro segmentaci slov. Podle tohoto dokumentu lze považovat každou syntaktickou kategorii (tj. podstatná jména, přídavná jména, číslovky, kvantifikátory, slovesa, příslovce ad.) za segmentovanou jednotku, tj. slovo (Liu et al 2013, s. 2).

Následný převod do pinyinu lze provést automaticky s poměrně vysokou (až s 95%) přesností pomocí operace vyhledávání z databáze. S chybovostí se lze setkat zejména u homonym (tj. jeden znak má více výslovností). Homonymům se jednoduše přiřazuje nejčastěji používaná výslovnost z korpusu (Liu a Guthrie 2009, s. 118; Peng et al. 2019, s. 5–6).

Jak však zmiňuje Qiao se svými kolegy (Qiao et al. 2010, s. 144), ani nejmodernější čínské automatické softwary pro segmentaci slov nejsou dostatečně výkonné a problém v jejich použití spočívá zejména v tom, že při segmentaci neznámých či nejednoznačných slov postupují nekonzistentně. Proto segmentace nebývá stoprocentně správná a musí projít dodatečnou kontrolou a opravou (Chen a Liu 2014, s. 83).

Další možností segmentace textu psaného čínskými znaky je k segmentaci přizvat rodilé mluvčí, kteří intuitivně dokážou odlišit hranice slov, což je však časově velmi náročná metoda, a navíc i tento způsob má svá úskalí, protože pravděpodobně ne všichni

Číňané rozliší hranice slov stejným způsobem, a proto by nebylo možné vytvořit jednotnou definici.

Další možností, kterou jsme využili i pro náš výzkum, je opřít se o jasně daná pravidla a text segmentovat manuálně. Díky dodržování pravidel by měl být výsledek segmentace vždy stejný bez ohledu na to, kdo segmentaci provádí. Pro náš výzkum jsme zvolili segmentovat jednotku *slovo* dle normy *Základní pravopisná pravidla pinyinu* GB/T 16159–2012 (která vychází ze starší verze normy GB/T 16159–1996). Normy vznikly na základě práce lingvistů a definují ortografická pravidla pro přepisování čínských znaků do abecedy pinyin a určují, kdy se mají slabiky psát dohromady a kdy odděleně. Tato pravidla a způsob dělení čínského textu by dle našeho názoru měla nejlépe vystihovat podstatu čínského slova, protože je to promyšlený systém s pevně danými pravidly.

Abychom ověřili hypotézu, že čínské ortografické slovo definované dle normy GB/T 16159–2012 sleduje ekonomická pravidla jazyka a odráží tak obecně platnou definici čínského slova, budeme respektovat návrh pana Wimmera a jeho spolupracovníků:

> „Keďže v každom jazyku možno použiť kritériá vedúce k odlišným výsledkom, odporúča sa na segmentáciu použiť buď Menzerathov zákon, alebo písanú formu slova. Menzerathov zákon má tú výhodu, že nie je konvenčný, je aplikovateľný na všetky jazyky bez ohľadu na ich písmo a dáva možnosť jednoznačného rozhodnutia sa pre tú segmentáciu, ktorá je s ním v najlepšej zhode“ (Wimmer 2003, s. 87).

2.3.3.3 Sporné případy při segmentaci dle normy GB/T 16159–2012

Při aplikování pravidel normy GB/T 16159–2012 jsme se setkali s případy, kdy byl převod do latinky nejednoznačný. Problematické bylo zejména určit, v jakých případech se píše podstatné jméno a záložka pro určení místa dohromady a kdy zvlášť. Dle autorů *Chinese Romanization: Pronunciation and Orthography* (1990) existuje mnoho příkladů tohoto typu konstrukcí, které se kvůli úzkým vztahům mezi jednotkami musí psát jako jeden celek (Yin a Felley 1990, s 125–129). Z toho důvodu jsme například následující spojení psali dohromady:

路上 lùshang　　街上 jiēshang

心里 xīnli　　地上 dìshang

身上 shēnshang

V souladu s normou jsme zapisovali dohromady i spojení ukazovacího slova zhè 这, nà 那 nebo nǎ 哪 spolu s bezprostředně následujícím znakem, a to v těchto případech:

这个 zhège　　那时 nàshí

这么 zhème　　那样 nàyàng

这样 zhèyàng　　那里 nàli

这时 zhèshí　　哪个 nǎge

那些 nàxiē

Mimoto jsme setkali se znaky, které se zpravidla píšou odděleně (např. znak zhǐ 只, yǒu 有), ale kniha *Chinese Romanization: Pronunciation and Orthography* uvádí konkrétní příklady, kdy se mají psát jako jedno slovo:

只要 zhǐyào　　有点 yǒudiǎn

只是 zhǐshì　　内外 nèiwài

只能 zhǐnéng　　前后 qiánhòu

只有 zhǐyǒu

Stejně tak kniha uvádí spojený zápis v případě použití postpozičních sloves, proto jsme je považovali také za jedno slovo:

走到 zǒudào　　出来 chūlai

回去 huíqu

Naopak odděleně jsme rozepisovali vazbu k vyjádření určení způsobu, tj. sloveso + příponu de 得 (případně bu 不) + adjektivum zvlášť. Například 说不清楚 a 看不清楚:

说 shuō　　看 kàn

不 bu　　不 bu

清楚 qīngchu　　清楚 qīngchu

Kromě výše zmíněných příkladů jsme se setkali s dalšími spornými výrazy, pro které nebylo definováno, jakým způsobem mají být zapsány. Proto jsme se rozhodli je členit následujícím způsobem:

得不到 débudào: píšeme dohromady z toho důvodu, že kniha uvádí spojený zápis pro slovo dédào

想不起来 xiǎngbuqǐlái: není definováno, zvolili jsme spojený zápis, tzn. spojení považujeme za jedno slovo

K ověření, jestli čínské ortografické slovo (vymezené dle normy GB/T 16159–2012) odpovídá ekonomizujícím pravidlům v jazyce, použijeme Menzerath-Altmannův zákon, který nám ukáže míru shody pro tento způsob segmentace.

Pro určení délky slova je za měřicí jednotku třeba zvolit slabiku. Pouze v tomto případě je možné zajistit obecnost, protože slabika může být použita ve všech jazycích (Popescu et al. 2013, s. 225). Další důvody, proč jsme za jazykovou jednotku na nižší jazykové úrovni zvolili právě slabiku, viz kapitola 2.3.2.

Kromě zvolené metody převedení čínských znaků do pinyinu dle normy GB/T 16159–2012, jsme za výběrové soubory zvolili i texty psané čínskými autory přímo v pinyinu. V tomto případě bylo na slova pohlíženo z ortografického hlediska – tedy jako souhrn grafémů mezi mezerami. Krátce se u těchto výběrových souborů ještě zastavme.

Jak již bylo uvedeno výše, autorka kapitoly *Mihuo* Zhang Liqing víceméně dodržuje rozdělení slov jako u normy GB/T 16159–2012 s výjimkou zápisů znaků 的, 得, 地 kdy znak 的 se přepisuje pouze jako d a znaky 得, 地 jako de. Další výjimky viz kapitola 2.2.4. V případě výskytu pomlčky, např. Dì-èr, zuǒ-yòu, shàng-xià, Ba-ba-ba jsme za jedno slovo považovali celé spojení.

U druhého výběrového souboru *Integrated Chinese Level 1 Part 2* bylo několik případů, kdy se rozdělení textu na slova odchylovalo od normy GB/T 16159–2012, např. slovesná (vido-časová) přípona guo 过 nebo znak le 了 byl vždy zapisován odděleně od slovesa, i když v normě je uveden stažený zápis (znak 了(le) by měl mít stažený zápis v případě, že se jedná o slovesnou příponu):

Lǐ Yǒu **qù guo** Táiběi hé Xiānggǎng, dànshì méi **qù guo** Běijīng,…

… **tīng le** Wáng Péng de jièshào hěn xiǎng qù Běijīng.

Tā **dìng le** kào zǒudào de wèizi, hái gěi Lǐ Yǒu dìng le yí fèn sùcān.

U obou výběrových souborů jsme respektovali původní způsob segmentace slov. Po aplikování statistických metod a vyhodnocení výsledků můžeme zjistit, zda tento způsob segmentace více vyhovuje ekonomizujícím pravidlům jazyka.

Slovo je bezprostřední složkou klauze.

2.3.4 Klauze (dānjù 单句)

Při segmentaci jsme dále zohlednili nejenom souvětí jako celek, ale i části, ze kterých se souvětí skládá, tzn. klauze. Klauze není jednoznačně definovaná jednotka a její vymezení má kvalitativní charakter, může tedy být definována více možnými způsoby. V případě vymezení klauze v čínštině se není možné opírat o grafický princip, jelikož použití interpunkce v čínské větě je značně nesystematické a liší se autor od autora. Avšak struktura čínské věty je víceméně pevně daná a větné členy jsou určeny velmi často právě svým postavením ve větě (Švarný 1993, s. 124):

> „Čínština, která nemá žádnou flexi ani aglutinaci, využívá pro vyjádření různých charakteristik slov a vztahů mezi slovy především slovosledu. … Slovosled, který je v čínštině syntakticky vázanější a zatíženější než slovosled ve flektivních i jiných jazycích s „volným“ slovosledem, dovoluje (případně i vyžaduje) přesuny větných členů, např. složek uvnitř vícenásobného přívlastku, jednotlivých složek při vícenásobném preverbálním a hlavně postverbálním určení predikativu apod., jen za určitých, gramatikou vymezených okolností.“

Při zohlednění tohoto faktu jsme se v případě čínských textů, jejichž věty mají pevný pořádek slov, rozhodli přistoupit k manuální segmentaci založené na gramatických vlastnostech čínštiny. Kromě toho jsme byli nuceni zohlednit i sémantické hledisko a kontext. Klauze v našem výzkumu představuje nejmenší samostatnou gramatickou jednotku a její vymezení vychází z následujících primárních pravidel:

1) Za klauzi považujeme struktury obsahující podmět a za ním přísudkové sloveso (případně adjektivum v postavení přísudku)
   a) Za podmětem a přísudkovým slovesem může následovat předmět (výjimečně na začátku věty)

b) dále klauze může obsahovat i příslovečná určení, která se kladou buď mezi podmět a přísudkové sloveso, nebo stojí v čele věty před podmětem, případně za slovesem;
c) přívlastek stojí vždy před slovem, který určuje;
d) záporka zpravidla stojí před přísudkovým slovesem;
e) tázací atonická částici ma 吗 stojí na konci věty, tázací větu je možné tvořit i zopakováním přísudkového slovesa, případně adjektiva (v tom případě se stále jedná o jednu klauzi)

2) za klauzi považujeme i struktury, ve kterých je podmět lexikálně nevyjádřený
3) klauze může být tvořena jedním základním větným členem. Základním větným členem může být jak sloveso v určitém nebo neurčitém tvaru, tak i výraz neslovesný (např. *Kdo?* Shéi? 谁？)
4) jedna klauze může obsahovat dvě za sebou následující slovesa (na prvním místě stojí modální sloveso, za kterým následuje další sloveso; např. *Také tam chci jít*. Wǒ yě xiǎng qù. 我也想去。)

Další zásady vycházejí z dostupných materiálů obsahující gramatická pravidla a konkrétní příklady jejich použití (Švarný 1993; Švarný 2001; Li et al. 2008).

Při identifikaci klauzí se není možné opírat o použití spojek, protože větné spojky souřadicí se v čínských větách vyskytují omezeně a jednotlivé klauze tak stojí vedle sebe. Stejně tak je tomu i u spojek pro věty předmětné, které v čínských větách chybí. Oproti tomu větné spojky podřadicí se ve větách běžně vyskytují, avšak fakultativně (Švarný 1993, s. 147; Švarný 2001, s. 150).

I když je použití interpunkce v čínské větě vágní, v mnohých případech může sloužit jako pomocný indikátor k segmentaci klauzí. Za klauze můžeme považovat části oddělené čárkami, případně středníkem, dvojtečkou, tečkou, vykřičníkem nebo otazníkem (ale také části, kde interpunkce zcela chybí). Při použití čárky je třeba zvážit, zda čárka odděluje jednotlivé klauze v souvětí, nebo větné členy, což je v čínštině běžný jev. Druhá varianta není pro účely naší segmentace považována za konec klauze.

Třebaže interpunkční znaménka nemají v čínských textech zažitá přesná pravidla, mohou být nápomocná při dělení souvětí na klauze. Proto se zaměříme na konkrétní případy jejich použití ve zkoumaných textech a upozorníme na zvláštní případy, se kterými jsme se setkali.

1) **Uvozovky**: v případě, že uvozovky vyznačovali přímou řeč, považovali jsme je za hranici klauze. V ostatních případech (zdůraznění, označení názvu) jsme je jako hranici klauze nezohlednili, např.

… 他**“呸”**的一声将牙签吐向桥下的河水，…[20]

… Tā **“pēi”** de yī shēng jiāng yáqiān tǔ xiàng qiáo xià de héshuǐ,…

他们**“呼哧呼哧”**的喘气声越来越重，…[21]

Tāmen “hūchihūchi” de chuǎnqì shēng yuèláiyuè zhòng,…

… 我把**“血”**字拉得又长又响，…[22]

… wǒ bǎ “xuè” zì lā de yòu cháng yòu xiǎng,…

…, 灯管拼凑出了**“黄昏咖啡馆”**这样五个字, …[23]

dēngguǎn pīncòu chū le “Huánghūn kāfēiguǎn” zhèyàng wǔ gè zì

…, yìsi shì “Bùyào”.[24]

2) **Čárka pro výčet**: čínština má kromě tradiční čárky také obrácenou čárku (、), která se zpravidla používá k oddělení souřadně spojených větných členů. Neslouží tedy k oddělení klauze. Dalším případem, kdy interpunkční znaménko neodděluje klauzi, je použití dlouhé pomlčky, za kterou navazuje výčet:

… 可是我总是同时回忆出四种牌子的香烟——前门、飞马、利群和西湖。[25]

… Kěshì wǒ zǒngshì tóngshí huíyìchū sì zhǒng páizi de xiāngyān——Qiánmén, Fēimǎ, Lìqún hé Xī-Hú.

[20] Vlastní překlad autorky: S„pch!“ vyplivl párátko do řeky pod mostem.
[21] Vlastní překlad autorky: Se zvukem „chu-či chu-či“ dýchali a funěli čím dál víc.
[22] Vlastní překlad autorky: Dlouze a hlasitě jsem vyslovil slovo „krev“.
[23] Vlastní překlad autorky: ..., neonová světla utvářela název z pěti znaků "Kavárna Soumrak".
[24] Příklad z textu zaznamenaného v pinyinu, proto nejsou uvedeny čínské znaky. Vlastní překlad autorky: ..., význam je „nechci“.
[25] Vlastní překlad autorky: ...ale vždycky si vzpomenu na čtyři značky cigaret: Qianmen, Feima, Liqun a Xihu.

3) **Určení místa a času**: tradičním jevem napříč všemi výběrovými soubory bylo oddělování příslovečných určení místa a času čárkou. Z našeho pohledu je nechápeme jako ukončení klauze. Uvádíme pouze některé příklady:

这天下午的时候，昆山走在大街上，…[26]

Zhè tiān xiàwǔ de shíhòu, Kūn Shān zǒu zài dàjiē shàng,…

然后，昆山向我们走来了，…[27]

Ránhòu, Kūn Shān xiàng wǒmen zǒu láile,…

Měitiān zhōngwǔ, suǒyǒu d háizi dōu děi shuìwǔjiào.[28]

4) **Arabská čísla**: v případě uvedení telefonního čísla pomocí arabských číslic byla před číslem vždy použita dvojtečka a po zápisu čísla vždy následovala tečka. Za ukončení klauze byla v těchto případech považována dvojtečka a číslo nebylo do segmentace zahrnuto. Např.

…，以后联系的电话改成：4014548。[29]

…, yǐhòu liánxì de diànhuà gǎichéng: 4014548.

…，号码是：8801946。[30]

…, hàomǎ shì: 8801946.

5) **Dvojtečka**: v jednom případě jsme za konec klauze nepovažovali dvojtečku, protože za ní následoval výčet. Celá věta byla považována za jednu klauzi:

…，她们的名字是：赵萍、张丽妮、沈宁。[31]

Tāmen de míngzi shì: Zhào Píng, Zhāng Lìnī, Shěn Nìng.

6) **Chybějící uvozovky:** V textu se vyskytlo mnoho vět, ve kterých chybělo oddělení klauzí čárkami nebo jinými interpunkčními znaménky. Nejčastější

[26] Vlastní překlad autorky: Odpoledne toho dne šel Kun Shan po ulici,…
[27] Vlastní překlad autorky: Potom k nám přišel Kun Shan …
[28] Vlastní překlad autorky: Každý den v poledne si všechny děti musely zdřímnout.
[29] Vlastní překlad autorky: ... telefonní číslo se poté změnilo na 4014548.
[30] Vlastní překlad autorky: ... číslo je: 8801946.
[31] Vlastní překlad autorky: ... jejich jména jsou Zhao Ping, Zhang Lini, Shen Ning.

důvod byl chybějící značení přímé řeči. I přes chybějící interpunkční znaménka jsme však klauze v segmentaci zohlednili:

…，**说**你问题太多，…[32]

…, shuō nǐ wèntí tài duō,…

我说我出门太急，…[33]

Wǒ shuō wǒ chūmén tài jí,…

朋友说这个问题不大，…[34]

Péngyou shuō zhège wèntí bú dà,…

Další možné pojetí klauze (měřené v průměrném počtu slov) v souvětí je určení na základě počtu sloves. Avšak mohou nastat i případy, kdy se v klauzi žádné sloveso nenachází, mělo by tedy platit $x = 0$. Pro naše výpočty jsme klauze bez sloves vždy započítali, nejkratší klauze má tedy minimální hodnotu 1, tj. $x \geq 1$.

2.3.5 Souvětí (fùjù 复句)

Jazykovou jednotku souvětí v této práci chápeme graficky jako úsek vymezený těmito interpunkčními znaménky: tečka 。 (jùhào, 句号), otazník ? (wènhào, 问号) a vykřičník ！(tànhào, 叹号). V několika případech jsme narazili na chybějící interpunkční znaménko značící konec přímé řeči a přímá řeč byla ukončena pouze uvozovkami nahoře (bez čárky či tečky). Logicky jsme je pak za konec souvětí považovali uvozovky nahoře ″ (yǐnhào, 引号). V případě výběrového souboru psaného přímo v pinyinu značilo začátek nového souvětí velké písmeno.

Z výše uvedeného vyplývá, že se nemusí nutně jednat o větné celky složené z více klauzí, ale pro účely segmentace za souvětí považujeme i větu holou, souvětí tak může být tvořeno jednou klauzí.

[32] Vlastní překlad autorky: ... řekl: „Máš příliš mnoho otázek“.
[33] Vlastní překlad autorky: ... řekl jsem: „Spěchám ven.“
[34] Vlastní překlad autorky: Kamarád mi řekl: „To není velký problém ...“

Jazykové úrovně

MAL stanoví, že všechny jazykové úrovně jsou navzájem propojené, konstrukt na jedné úrovni může zároveň být konstituentem na jiné úrovni. V našem případě spojením výše uvedených jednotek do vzájemných vtahů získáme tři jazykové úrovně ($U_i, i = 1, 2, 3$). Na nejnižší jazykové úrovni U3 je konstruktem x_3 slovo (měřené ve slabikách) a konstituentem y_3 je slabika (měřená v průměrném počtu grafémů). Na jazykové úrovni U2 je konstruktem x_2 klauze (měřená ve slovech) a konstituentem y_2 slovo (měřené v průměrném počtu slabik). Na nejvyšší jazykové úrovni U1 je konstruktem x_1 souvětí (měřené v klauzích) a konstituentem y_1 klauze (měřená v průměrném počtu slov).

Pro přehlednost jsou definované binarismy uvedeny v Tabulce 11.

Tabulka 11 Jazykové úrovně *Ui*, x_i konstrukt, y_i konstituent (i=1, 2, 3)

Jazyková úroveň	**Konstrukt x_i; konstituent y_i**		**Délka**
U3	x_3	slovo	ve slabikách
	y_3	slabika	v průměrném počtu grafémů
U2	x_2	klauze	ve slovech
	y_2	slovo	v průměrném počtu slabik
U1	x_1	souvětí	v klauzích
	y_1	klauze	v průměrném počtu slov

Největší pozornost budeme věnovat jazykové úrovni U3, na které se slovo nachází na pozici konstruktu. Třebaže je tato jazyková úroveň pro nás klíčová, zákon by se samozřejmě měl odrážet na všech jazykových hladinách. Na první pohled by se mohlo zdát, že by bylo uspokojivé zkoumat pouze ty jazykové úrovně, na kterých *slovo* vystupuje jako konstrukt či konstituent. Z našeho hlediska by to však bylo nedostačující, protože jazyková jednotka *slovo* ovlivňuje výsledky i na nejvyšší jazykové úrovni U1 souvětí – klauze, kde slouží jako jednotka pro výpočet průměrných délek klauzí, proto zkoumáme právě tyto tři jazykové úrovně.

2.4 Segmentace a kvantifikace

Po zvolení vhodných výběrových souborů, určení jazykových jednotek a jazykových úrovní můžeme přistoupit k segmentaci výběrových souborů. Při jejich segmentaci jsme se rozhodli vynechat nadpisy, podnadpisy, data vydání, abstrakty,

klíčová slova, poznámky pod čarou či bibliografii, abychom získali co nejsouvislejší text. Prostředkem pro ověření vzájemných vztahů na jazykových úrovních bude MAL.

Za dodržení výše uvedených pravidel jsme text segmentovali nejprve za použití programu Wenlin 4 převedením znaků pomocí funkce *Make transformed copy – Pinyin transcription* a následně jsme každý převod zkontrolovali a opravili dle normy GB/T 16159–2012. Konkrétně jsme postupovali následujícím způsobem:

Příklad segmentace

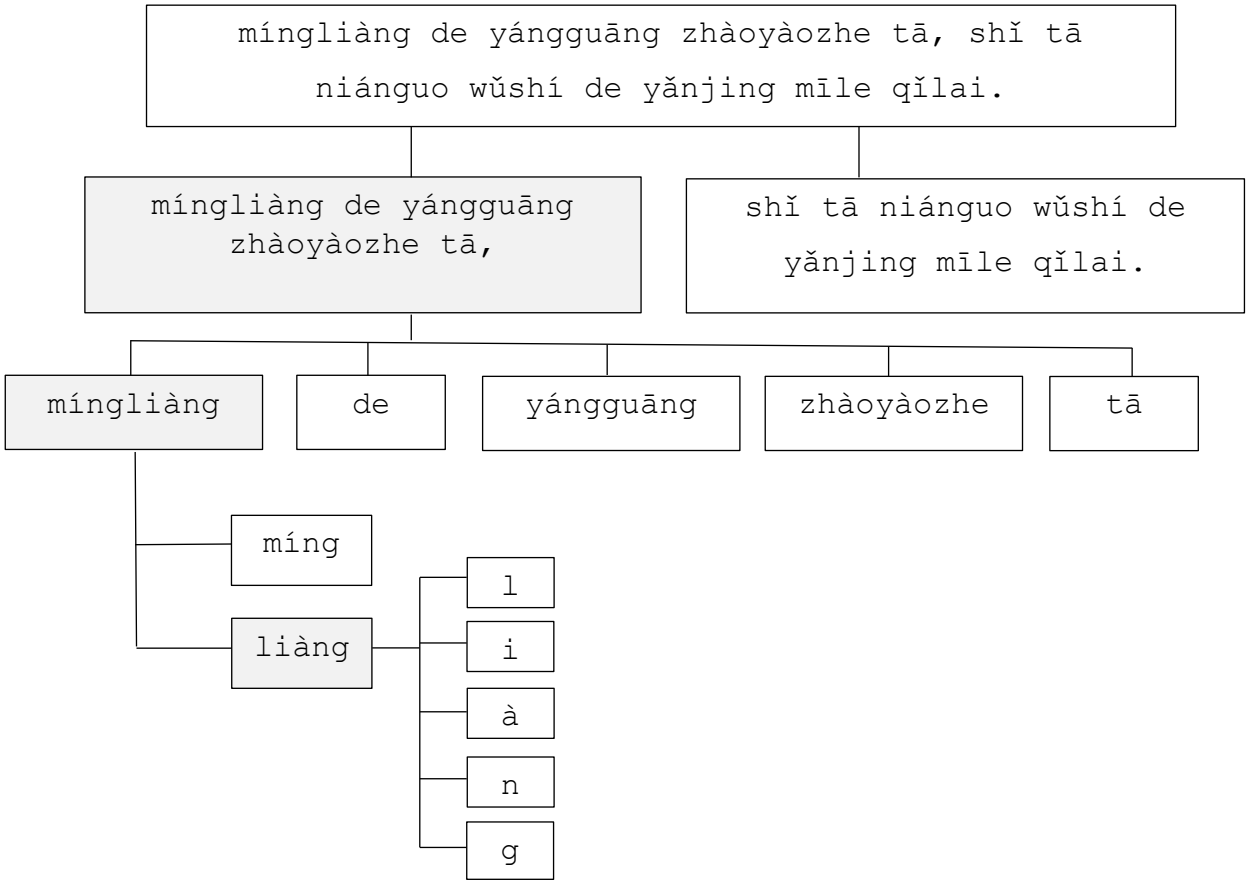
1) Původní text ve znacích:

明亮的阳光照耀着他，使他年过五十的眼睛眯了起来。(Yu 2011, s. 52)

2) Text převedený do abecedy pinyin podle pravidel uvedených v normě *Basic rules of the Chinese phonetic alphabet orthography* (GB/T 16159–2012):

míngliàng de yángguāng zhàoyàozhe tā, shǐ tā niánguo wǔshí de yǎnjing mī le qǐlai.

3) Po převedení textu do abecedy pinyin by segmentace na jednotlivých jazykových úrovních vypadala následovně:



Obrázek 9 Příklad segmentace

Obrázek 9 nejprve znázorňuje rozklad jednoho souvětí na dvě klauze. Segmentace pokračuje rozčleněním klauzí na slova, která jsou poté segmentována na slabiky. Na nejnižší úrovni jsou pak slabiky rozsegmentovány na grafémy.

Po provedení segmentace budeme text kvantifikovat. Ze získaných dat sestavíme tabulky pro hodnoty konstruktů (x_i, i=1, 2, 3), spolu s jejich četnostmi (z_i, i=1, 2, 3) a délky konstituentů na jednotlivých uvažovaných jazykových úrovních (y_i, i=1, 2, 3). Segmentaci a následnou kvantifikaci dat budeme realizovat v programu Microsoft Excel, který se z předchozích zkušeností jeví jako nejvhodnější nástroj.

2.5 Testování spolehlivosti modelu pomocí statistických metod

Hodnoty v tabulkách, které získáme po kvantifikaci textů, budou sloužit jako vstupní data pro výpočet parametrů A a b. K ověření spolehlivosti modelu MAL, tedy k verifikaci, jak těsně přiléhá regresní křivka k izolovaným bodům znázorňující naše pozorování, použijeme koeficient determinace R^2. Interval koeficientu determinace R^2 se pohybuje v rozmezí 0–1. Čím více se hodnota blíží k 1, tím lépe model sedí. „Hodnoty R^2 větší nebo rovny 0,7 mohou prokazovat adekvátní a dobře sedící model v kvantitativní lingvistice“ (Benešová 2011, s. 47, 77). Vzhledem k přirozenému jazyku je hodnota vyšší než 0,7 v lingvistickém experimentu považována za dostačující[35].

Pro výpočet parametrů A a b a dalších statistických údajů jsme zvolili software MA Studio, který byl vytvořen na Katedře obecné lingvistiky Univerzity Palackého v Olomouci. Software umožňuje použití několika různých formulí MALu s eventuálním využitím vah, které odráží četnosti jednotlivých jednotek. V našem výzkumu budeme používat zkrácenou verzi $\boldsymbol{y = Ax^{-b}}$ (Model 2), viz kapitola 2.

Z důvodu využití vah jsme do výpočtů zahrnuli i nejméně frekventovaná pozorování, tj. extrémy. V případě, že bychom se rozhodli pro jejich vynechání, budeme na tyto konkrétní případy vždy upozorňovat.

[35] Ústní diskuse s profesorem Prof. Ing. RNDr. Lubomír Kubáčkem, DrSc., Dr.h.c. z Katedry matematické analýzy a aplikací matematiky, Přírodovědecká fakulta UPOL.

3 Interpretace získaných dat

Na následujících stránkách podrobně rozebereme a interpretujeme veškerá data získaná z kvantitativní analýzy všech výběrových souborů a budeme se zamýšlet nad možnými příčinami ať už výsledků, které by mohly potvrzovat naši hypotézu, tak i těch výsledků, které by ji mohly vyvracet.

U každého výběrového souboru na dané jazykové úrovni U*i* nejprve uvedeme tabulku vztahující se ke konkrétním výpočtům konstruktů x_i, jejich četnosti z_i a délky konstituentů y_i, které jsme získali z programu Microsoft Excel (se zaokrouhlením na dvě desetinná místa), jazyková úroveň $i = 1, 2, 3$. Vzhledem k použité metodě ověřování jsme do výpočtů zahrnuli veškerá data, tzn. tokeny, protože nás zajímají konkrétní texty, a ne pouze jejich části. Teoretickou reprezentaci, tedy typy, blíže nezkoumáme. Za tabulkou s výpočty následují obrázky s grafickým zobrazením křivek, které ilustrují tendenci MALu, model 2. Křivky vygenerované pomocí programu MA Studio nekopírují přesně tendenci zadaných bodů, ale vzhledem k zadaným matematickým vzorcům se snaží pomocí metody nejmenších čtverců křivkou co nejlépe vystihnout odpovídající tendenci.

Prostřednictvím softwaru MA Studio dále získáme výpočty parametrů A a b a koeficientu determinace R^2, jejichž hodnoty přehledně uvedeme v tabulce následující po grafických vizualizacích.

Při interpretaci dat budeme stejně jako při segmentaci postupovat nejprve od nejnižšího jazykového binarismu slovo – slabika na úrovni U3 k nejvyššímu jazykovému binarismu souvětí – klauze na úrovni U1.

3.1 Jazyková úroveň U3: slovo – slabika

Na této jazykové úrovni budeme zkoumat vzájemný vztah dvou jazykových jednotek slova a slabiky, slova coby konstruktu x_3 měřeného v počtu slabik a slabiky jako jeho konstituentu y_3 měřeného v průměrném počtu grafémů, z_3 udává frekvenci slov dané délky. Budeme zjišťovat, jestli způsob převodu čínských znaků do pinyinu dle normy GB/T 16159–2012 sleduje ekonomická pravidla jazyka a odráží tak obecnou definici čínského slova. Jako nástroj ověřování nám bude sloužit MAL. Nejnižší jazyková úroveň U3 je pro nás klíčová, proto jí budeme věnovat nejvíce pozornosti, ale zákon by se

samozřejmě měl odrážet na všech jazykových úrovních. Předmětem zkoumání budou všechny zvolené výběrové soubory 1–5.

Na nižších jazykových úrovních bývá zpravidla dostatečné množství vstupních dat, což je i náš případ. Konkrétní data získaná kvantifikací všech výběrových souborů uvádíme v následující Tabulce 12.

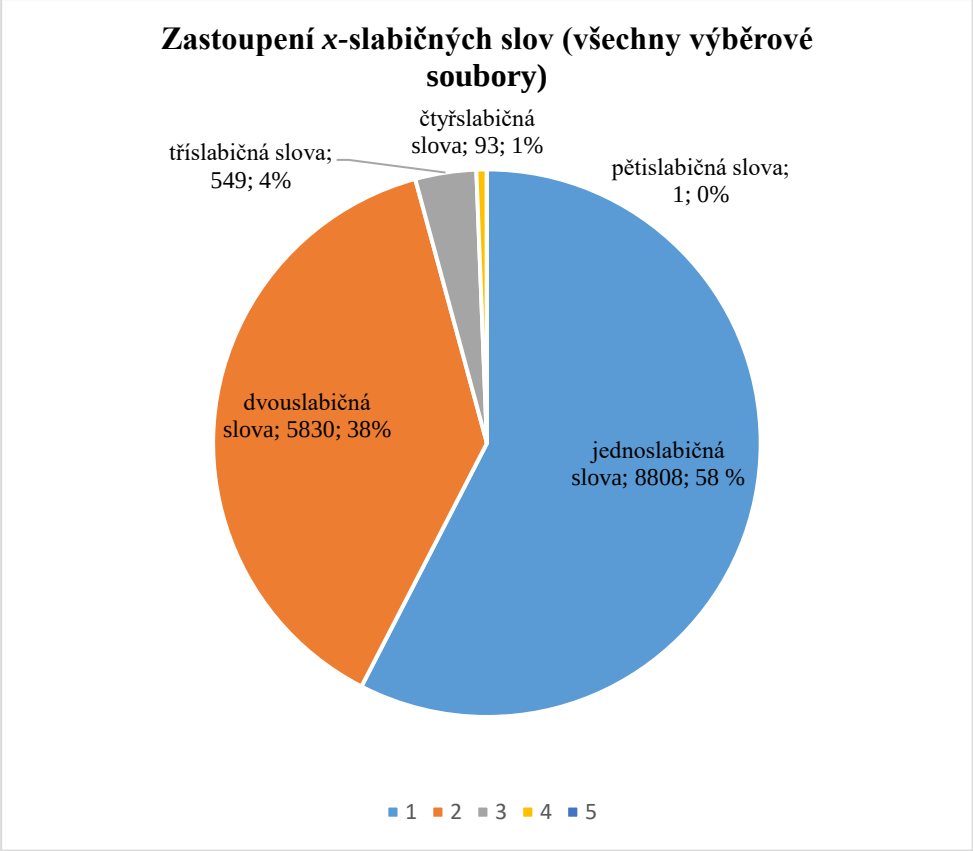
Tabulka 12 Jazyková úroveň U3: slovo (měřené v slabikách) – slabika (měřená v průměrném počtu grafémů)

Výběrový soubor	**Délka slova v slabikách x_3**	**Frekvence z_3**	**Průměrná délka slabik v grafémech y_3**
Výběrový soubor 1	1	2 686	2,92
	2	1 620	3,03
	3	234	2,96
	4	31	2,99
Výběrový soubor 2	1	3 446	2,76
	2	2 194	3,02
	3	166	2,83
	4	47	2,96
	5	1	2,40
Výběrový soubor 3	1	925	2,73
	2	668	3,04
	3	21	3,10
	4	10	3,08
Výběrový soubor 4	1	1 431	2,58
	2	1 210	2,98
	3	125	3,01
	4	5	2,70
Výběrový soubor 5	1	320	3,01
	2	138	3,06
	3	3	3,00

Z Tabulky 12 je patrné, že po převodu čínských textů do latinky dle normy GB/T 16159–2012 se ve výběrových souborech vyskytla slova s délkou jeden až pět znaků, z nichž nejčetnější jsou u každého výběrového souboru slova jednoslabičná a za nimi následují slova dvouslabičná. Slova víceslabičná mají procentuálně mnohem nižší zastoupení (viz Tabulka 13 a Obrázek 10). Výjimkou je výskyt pětislabičného slova (*obávat se* tǎntèbùānzhe 忐忑不安着) ve Výběrovém souboru 2, kde se však pětislabičné slovo objevilo pouze jednou, pro výzkum má tedy zanedbatelný význam.

Tabulka 13 Zastoupení *x*-slabičných slov ve všech výběrových souborech bez odebrání duplicit

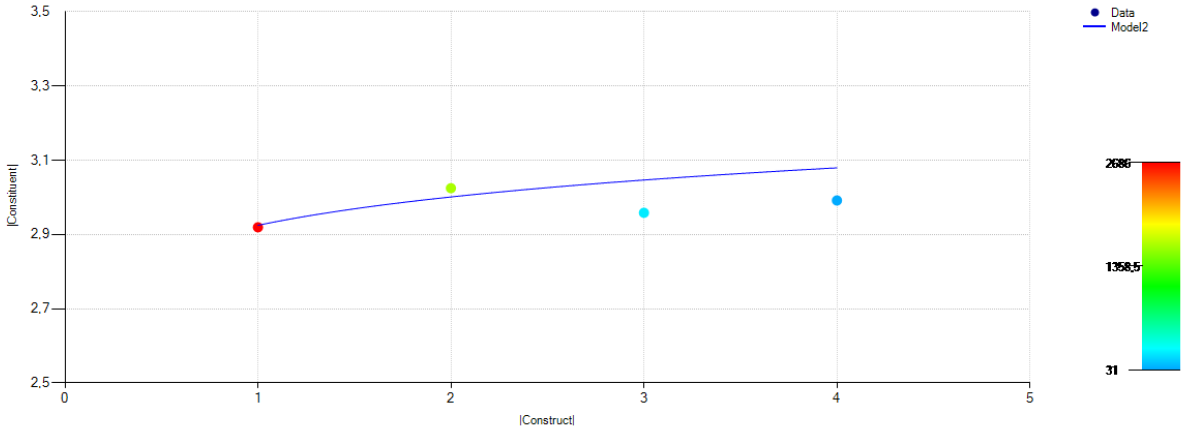
Počet slabik	Počet výskytů	Procentuální zastoupení
1	8 808	57,64 %
2	5 830	38,15 %
3	549	3,59 %
4	93	0,61 %
5	1	0,01 %



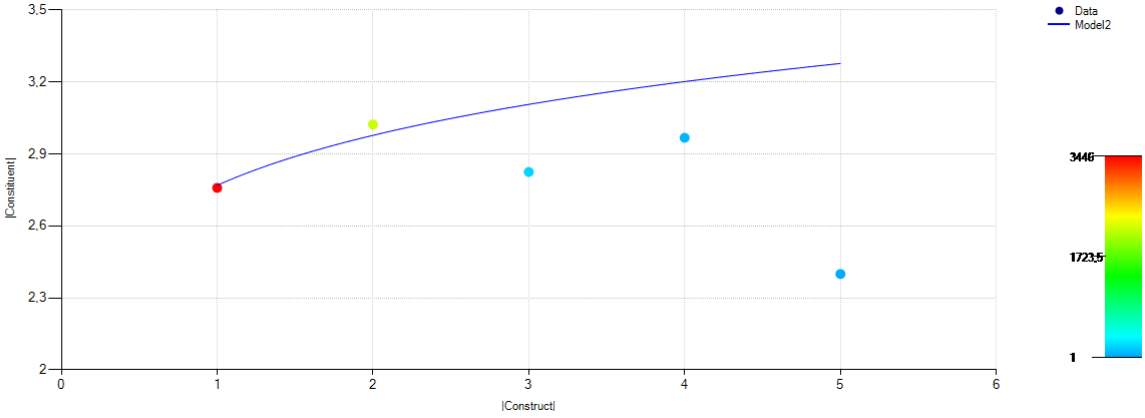
Obrázek 10 Zastoupení *x*-slabičných slov ve všech výběrových souborech bez odebrání duplicit

Pokud se zaměříme na délky slabik v grafémech, zjistíme, že se jejich průměrné hodnoty pohybují v intervalu ⟨2,40; 3,10⟩, tedy o délce intervalu 0,70. Zajímavým zjištěním je, že s výjimkou Výběrového souboru 5 je průměrná délka slabik jednoslabičných slov kratší než průměrná dálka slabik dvouslabičných, tříslabičných i čtyřslabičných slov, což je v rozporu s tendencí MALu a také s ekonomizací jazyka.

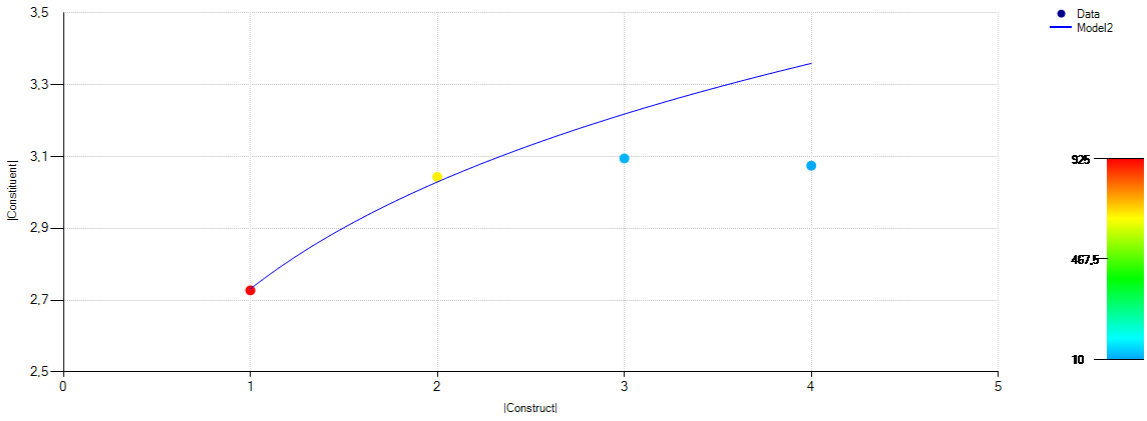
Pro lepší ilustraci získaných dat z Tabulky 12 uvádíme grafické vizualizace jednotlivých výběrových souborů, které zachycují modely MALu s implementovanými vahami, což je dle autorů článku Optimization of Parameters in the Menzerath–Altmann Law, II (Andres et al. 2014) považováno za vhodný model.



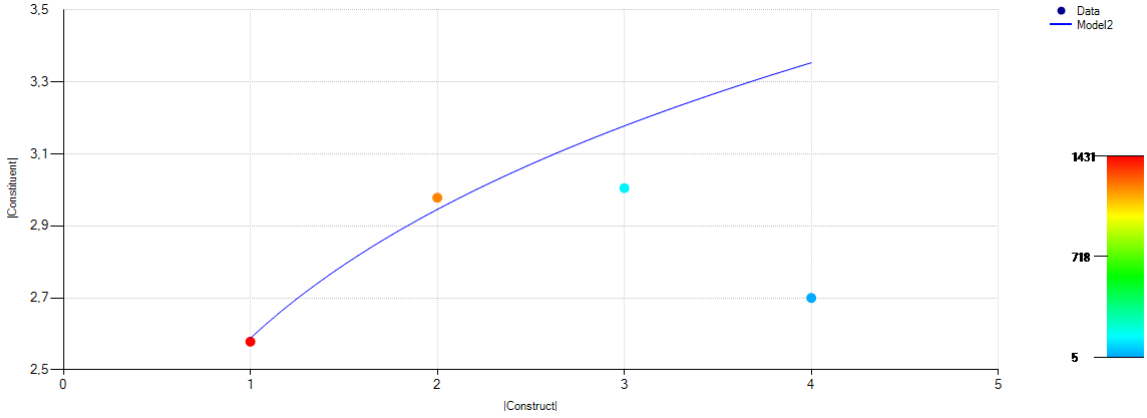
Obrázek 11 Grafická vizualizace pozorování uvedených v Tabulce 12 s proloženou regresní křivkou, Výběrový soubor 1: Yu Hua – *Kamarádi* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů)



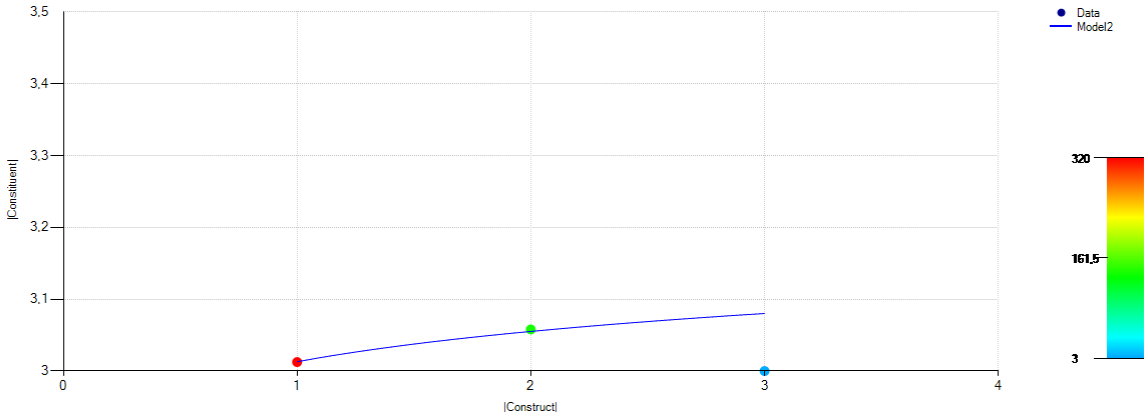
Obrázek 12 Grafická vizualizace pozorování uvedených v Tabulce 12 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů)



Obrázek 13 Grafická vizualizace pozorování uvedených v Tabulce 12 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů)



Obrázek 14 Grafická vizualizace pozorování uvedených v Tabulce 12 s proloženou regresní křivkou, Výběrový soubor 4: Zhang Liqing – *Deníkové záznamy v pinyinu* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů)



Obrázek 15 Grafická vizualizace pozorování uvedených v Tabulce12 s proloženou regresní křivkou, Výběrový soubor 5: *Integrated Chinese Level 1 Part 2* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů)

Z výše uvedených grafů je na první pohled zřejmé, že žádný výběrový soubor nevykazuje tendenci MALu. Namísto požadované konvexní křivky, která ilustruje grafickou podobu MALu, je ve všech případech průběh křivek konkávní. Křivky jsou navíc rostoucí, což je dáno zápornými hodnotami parametrů *b*. Pro úplnost uvádíme také hodnoty koeficientů determinace R^2, které sice vykazují i vysoké hodnoty, ale tyto hodnoty jsou platné pro stoupající tendenci křivek, což je v rozporu s MALem. Jistě je zajímavé, že všechny výběrové soubory kopírují totožné tendence. Konkrétní hodnoty parametrů a koeficientů determinace viz Tabulka 14. Normalita a homoskedasticita jsou splněny.

Tabulka 14 Parametry *A* a *b* a koeficient determinace R^2 pro matematický model 2 MALu vztahující se k empiricky získaným datům na úrovni U3

Výběrový soubor	**Parametr *A***	**Parametr *b***	***Koeficient determinace* R^2**
Výběrový soubor 1	2,93	-0,03	0,74
Výběrový soubor 2	2,77	-0,10	0,79
Výběrový soubor 3	2,73	-0,15	0,97
Výběrový soubor 4	2,59	-0,19	0,95
Výběrový soubor 5	3,01	-0,02	0,90

Prvním možným vysvětlením, proč text převedený z čínských znaků do latinky neodpovídá na nejnižší úrovni U3 předpokladu MALu, je, že v abecedě pinyin se vyskytují spřežky a dva grafémy označují jediný zvuk. Příkladem může být slabika `chuang`, která je zapsána pomocí šesti grafémů, avšak iniciální přídechová souhláska `ch` se vysloví jako jeden zvuk [tʂʰ] a zadní nosovka `-ng` jako [ŋ], celkem se tedy jedná o čtyři hlásky.

Druhým vysvětlením může být, že způsob segmentace řídící se normou GB/T 16159–2012 není tím nejvhodnějším způsobem členění čínského textu na slova. Dle zmíněného způsobu segmentace získáme jako nejčetnější skupinu jednoslabičná slova, což by se na první pohled mohlo zdát, že je v rozporu se současným stavem nejpoužívanější čínské slovní zásoby. Čínština se ve svém 3 500letém vývoji postupně měnila z monosylabického jazyka na jazyk používající víceslabičné morfémy. Dle vědce

Lü je 61 % z 3 000 nejpoužívanějších čínských slov z čínské slovní zásoby dvouslabičných (jak cituje Sun 2006, s. 49–50). Stejné tendence v čínském jazyce vysledoval i lingvista Wang, který prováděl studie délky slova na statickém a dynamickém korpusu zahrnujícím texty o více než milionu slov a dokonce tvrdí, že nejenom frekvence dvouslabičných slov ze statického korpusu (typy), ale i frekvence dvouslabičných slov z dynamického korpusu (tokeny) převyšují frekvenci jednoslabičných slov (Wang 2013, s. 39–42). Dalšími čínskými vědci zabývajícími se studiem délky slov jsou Chen, Liu a Liang, kteří sledovali vývoj délek čínských slov v historickém vývoji. I dle jejich závěru se v čínském jazyce projevuje trend používání víceslabičných slov a ověřili, že frekvence víceslabičných slov neustále roste. Avšak data získaná studiem statického a dynamického korpusu se v jejich studii liší. V případě statického korpusu jsou nefrekventovanější dvouslabičná slova, zatímco u dynamického korpusu jsou to slova jednoslabičná (Chen et al. 2014; Chen et al. 2015).

Při výzkumu frekvence čínských slov je tedy nutné rozlišovat, zda zkoumáme statické či dynamické soubory, tj. typy nebo tokeny, protože jejich frekvence se markantně liší. Pokud budeme zohledňovat pouze typy (lexikální jednotky čínské slovní zásoby, tj. slovníkové vstupy), zde je převaha dvouslabičných slov (přibližně 74 % oproti 12 % jednoslabičných slov, zbytek jsou slova tříslabičná, čtyřslabičná a pětislabičná). Ale pokud zkoumáme tokeny (lexikální jednotky použité v textu), většinu tvoří jednoslabičná slova (přibližně 64 % oproti 34 % dvouslabičných slov; Breiter 1994, s. 230; Třísková 2010, s. 48). V případě našich výběrových souborů je to přibližně 58 %. Důvodem je použití krátkých jednoslabičných slov, převážně funkčních, která jsou v textech velmi frekventovaná.

Tento fakt nás vede k formulaci hypotézy, že většina jednoslabičných slov jsou i v případě našich výběrových souborů funkční slova, která se v textech často opakují, a právě funkční slova pravděpodobně způsobují rostoucí tendenci, která je v rozporu s tendencí MALu. Tento předpoklad je v souladu s jazykovým zákonem o stručnosti (Brevity law, též nazývaný jako Zipfův zákon o zkracování), který tvrdí, že čím je slovo frekventovanější, tím bývá kratší a naopak, čím je slovo méně frekventované, tím bývá delší (Bentz a Ferrer-i-Cancho 2016). Proto jsme se rozhodli se na jednoslabičná slova více zaměřit a zjistit o jaké skupiny slov se jedná a případně se rozhodnout pro jiný typ segmentace textu na slova, díky kterému bychom dostali data, která budou lépe odpovídat předpokladům MALu. V následujícím experimentu budeme porovnávat výše uvedené výsledky s výsledky, které získáme na základě alternativního způsobu segmentace.

3.1.1 Experiment 1

V následující části se zaměříme na převádění textu psaného v čínských znacích do pinyinu a budeme tedy používat pouze data ze třech výběrových souborů, tj. Výběrový soubor 1, 2 a 3. Pouze tyto výběrové soubory byly zaznamenány pomocí čínského znakového písma. Ostatní dva výběrové soubory již byly zaznamenány pinyinem, který nebudeme upravovat, protože tento způsob segmentace slouží zejména k účelům porovnání. U těchto dvou výběrových souborů akceptujeme autorské členění textu. Výzkum je zaměřený pouze na tokeny.

Dle MALu by mělo platit, že jednoslabičná slova měřená v průměrném počtu slabik by měla být nejdelší. Proto je předpokladem, že klesající tendenci narušují jednoslabičná slova krátké délky, tj. slova složená z malého počtu grafémů. Z toho důvodu nebudeme provádět segmentaci vycházející ze způsobu členění Výběrového souboru 5, protože ten rozděloval slova na ještě kratší celky a získali bychom ještě více krátkých jednoslabičných slov.

Při podrobnějším zkoumání jednoslabičných slov jsme zjistili, že v rámci každého výběrového souboru je prvních 10 nejčetnějších slov vždy jednoslabičných a nejčetnějším slovem ve všech třech výběrových souborech bylo přivlastňovací slovce de 的 (Výběrový soubor 1: 6,91 %, Výběrový soubor 2: 5,46 % a Výběrový soubor 3: 5,47 % ze všech slov) následované osobními zájmeny *já* wǒ 我 (Výběrový soubor 1 a Výběrový soubor 3), případně *on/ona* tā 他/她 (Výběrový soubor 2), dále viz Tabulka 15.

Tabulka 15 Prvních 10 nejčetnějších slov v rámci Výběrových souborů 1–3

	Výběrový soubor 1			**Výběrový soubor 2**			**Výběrový soubor 3**		
Rank	**Znaky**	**Pinyin**	**Frekvence**	**Znaky**	**Pinyin**	**Frekvence**	**Znaky**	**Pinyin**	**Frekvence**
1	的	de	316	的	de	316	的	de	89
2	我	wǒ	188	她	tā	268	我	wǒ	79
3	他	tā	150	他	tā	242	你	nǐ	50
4	昆	Kūn	122	在	zài	138	不	bu	42
5	山	Shān	122	了	le	107	是	shì	29
6	刚	Gāng	90	一	yī	94	一	yī	26
7	石	Shí	90	我	wǒ	93	说	shuō	22
8	一	yi	85	红	Hóng	90	个	ge	20

9	在	zài	65	林	Lín	89	就	jiù	18
10	上	shàng	59	不	bù	87	在	zài	18

Na základě dat z Tabulky 15 nás zajímá zejména výskyt nejfrekventovanějšího slovce de 的, protože se jednak jedná o funkční slovo a jednak jeho četnost je tak vysoká, že jistě může ovlivňovat celkové výsledky v rámci každého výběrového souboru. Je možné, že právě jeho použití jako samostatného slova by mohlo velkou mírou přispívat k obrácené tendenci, která je u všech třech výběrových souborů v rozporu s tendencí MALu. Podívejme se tedy znovu na normu a ověřme správnost přepisu. Jak již bylo zmíněno dříve, i samotná norma povoluje více možných forem převodu znaků do latinky. Konkrétně bod 6.1.9.1 se týká zmíněného znaku de 的 a dalších znaků de 地 a de 得, u kterých norma prvotně definuje zapisovat slova s mezerou, ale kromě toho také umožňuje stažený zápis v těch případech, kdy před zmíněnými znaky stojí jednoslabičné slovo. Kromě tohoto bodu můžeme v normě narazit i na další odchylky – například stejná možnost dvojího zápisu je uvedena u znaků zhè 这, nà 那 a nǎ 哪 (bod 6.1.4.2). U těchto znaků norma definuje přepisovat znaky do pinyinu samostatně, ale poté vyjmenovává výjimky, kdy je možné znaky přepisovat společně s následujícím znakem jako jedno slovo. Další skupinou slov, která jsou v normě definována dvojím způsobem, jsou záložky (bod 6.1.1.1 a 6.1.1.2). Norma uvádí, že v případě, že spojení se záložkou tvoří ustálený výraz, píšou se dohromady. V opačném případě zvlášť. Dále norma umožňuje v některých případech stažený zápis deiktického slova a numerativu.

V následujícím experimentu budeme vycházet z těchto bodů, a kromě toho také zohledníme pravidla segmentování, která vychází z Výběrového souboru 4, který v některých případech spojoval záložky a záporky se slovy plnovýznamovými. Zjištěná pravidla ještě více zobecníme a v našem experimentu budeme stažený zápis aplikovat nejenom pro uvedené výjimky, ale pro všechny případy, ve kterých norma umožňuje variantní formy zápisu. Tzn. pokud norma definuje, že například stažený tvar je možné zvolit pouze v případě, že znaku předchází jednoslabičné slovo, my budeme stažený tvar volit i pro víceslabičná slova apod.

Po opětovném prostudování normy a zohlednění pravidel segmentování Výběrového souboru 4, jsme mezi stažený způsob zápisu v latince doplnili následující případy:

1) připojení **záložek** k podstatnému jménu: konkrétně se jedná o shàng 上, xià 下, lǐ 里, zhōng 中, biān 边, miàn 面, qián 前, hòu 后, páng 旁, biān 边
2) připojení **atributivního slovce de 的** k podstatnému jménu či zájmenu pro účely vytvoření přívlastku
3) připojení **pomocného ukazatele de 地** k adjektivu pro účely vytvoření určení způsobu
4) připojení **ukazatele de 得** ke slovesu či adjektivu pro účely vytvoření komplementu
5) připojení **předpony dì 第** pro účely vytvoření řadových číslovek (v normě oddělováno pomlčkou)
6) připojení **numerativu k číslovce**[36] (dle vzoru nǎge 哪个)
7) připojení **postpozičních sloves zài 在 a dào 到** k předcházejícímu slovesu
8) připojení **záporek bù 不 a méi 没** (dle vzoru méiyǒu 没有, které se píše dohromady)

Po vymezení námi zvolených pravidel jsme na výběrové soubory aplikovali daná pravidla a znovu jsme je segmentovali. Získané výsledky následují v Tabulce 16.

Tabulka 16 Experiment 1: Jazyková úroveň U3: slovo (měřené v slabikách) – slabika (měřená v průměrném počtu grafémů) pro námi navržený způsob segmentace

Výběrový soubor	**Délka slova v slabikách x_3**	**Frekvence z_3**	**Průměrná délka slabik v grafémech y_3**
Výběrový soubor 1	1	1 692	3,06
	2	1 674	2,99
	3	458	2,85
	4	62	3,01

[36] Je nutné odlišit numerativy od měrových jednotek a měrových jmen, které jsou plnovýznamové a stojí tak samostatně. I když bychom zjednodušeně řečeno mohli numerativy chápat jako neurčité členy v angličtině, není možné je připojit k substantivu, a to z toho důvodu, že mezi numerativ a podstatné jméno je možné vkládat adjektivum, např. yībǎ liànghuǎnghuǎng de càidāo 一把亮晃晃的菜刀.

	5	18	2,80
Výběrový soubor 2	1	2 275	2,82
	2	2 208	2,97
	3	487	2,82
	4	55	2,93
	5	12	2,65
Výběrový soubor 3	1	631	2,82
	2	689	2,99
	3	97	2,82
	4	13	3,00
	5	2	2,60

Dle získaných dat uvedených v Tabulce 16 se oproti původnímu výzkumu podstatně navýšily hodnoty víceslabičných slov, zejména dvouslabičných a tříslabičných, ale můžeme pozorovat i významný nárůst například u pětislabičných slov, u kterých se počet pozorování navýšil z jednoho dokonce na 32.

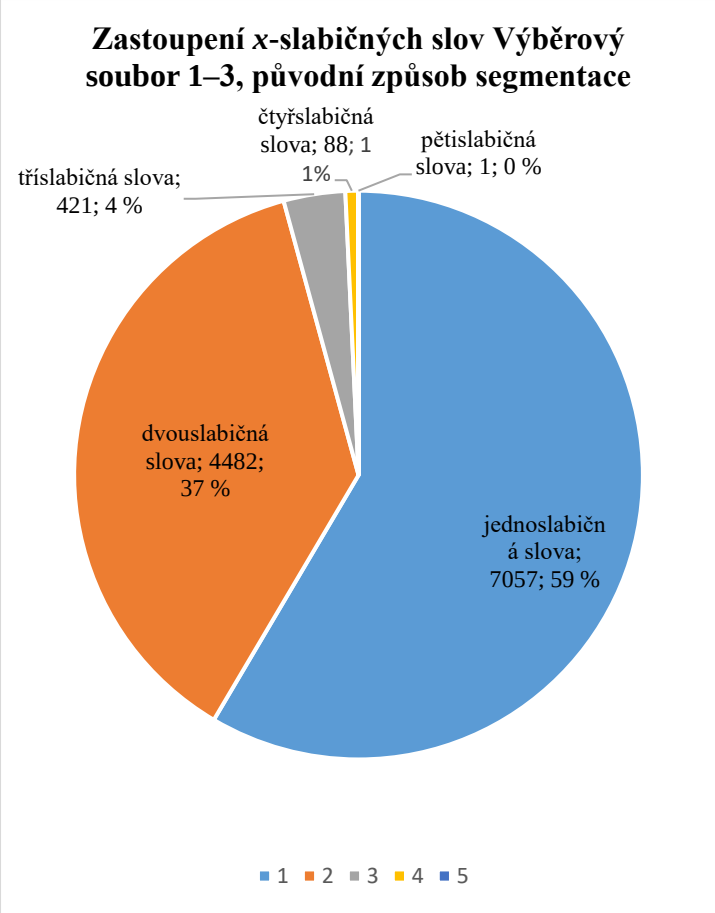
Pro porovnání připojujeme tabulky a grafy zastoupení *x*-slabičných slov u Výběrových souborů 1–3 pro oba způsoby segmentace, viz Tabulka 17 a 18, Obrázek 16 a 17.

Tabulka 17 Zastoupení *x*-slabičných slov, Výběrový soubor 1–3, původní způsob segmentace dle normy GB/T 16159–2012

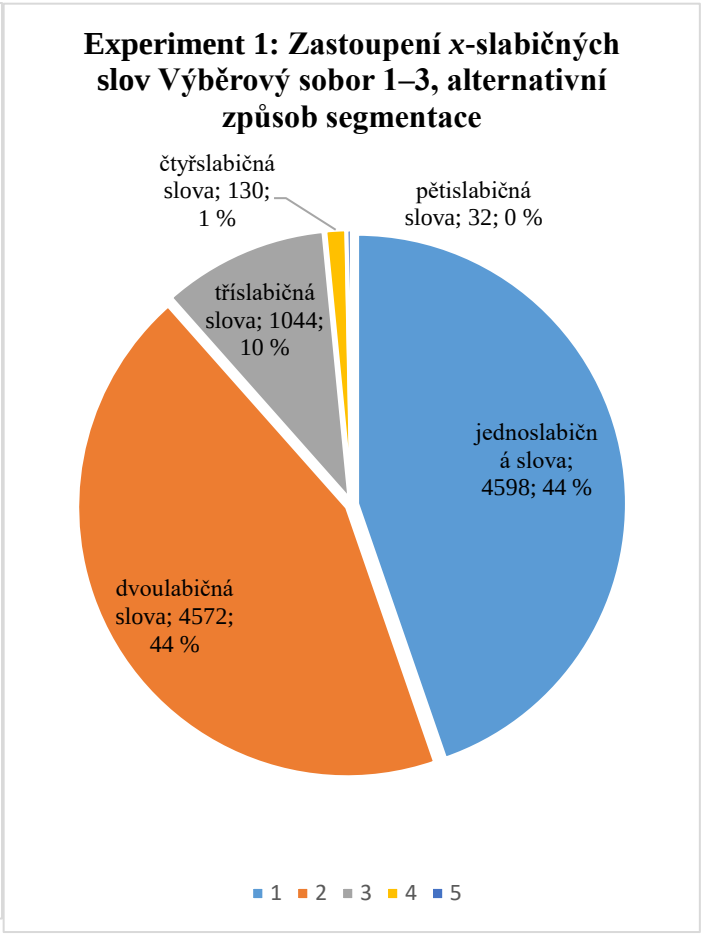
Celkem Výběrový soubor 1–3	1	7 057	58,57 %
	2	4 482	37,20 %
	3	421	3,49 %
	4	88	0,73 %
	5	1	0,01 %

Tabulka 18 Experiment 1: Zastoupení *x*-slabičných slov, Výběrový soubor 1–3, alternativní způsob segmentace

Celkem Výběrový soubor 1–3	1	4 598	44,31 %
	2	4 572	44,06 %
	3	1 044	10,06 %
	4	130	1,25 %
	5	32	0,31 %

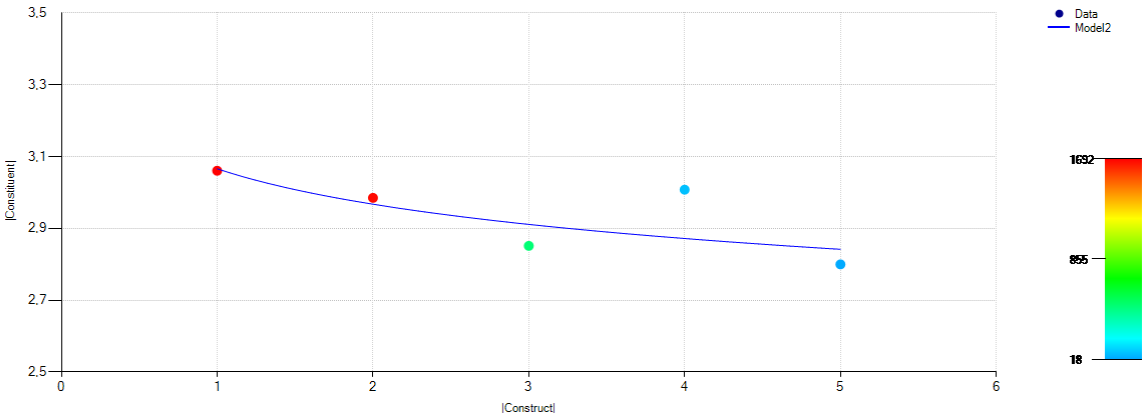


Obrázek 16 Zastoupení *x*-slabičných slov, Výběrové soubory 1–3, původní způsob segmentace dle normy GB/T 16159–2012

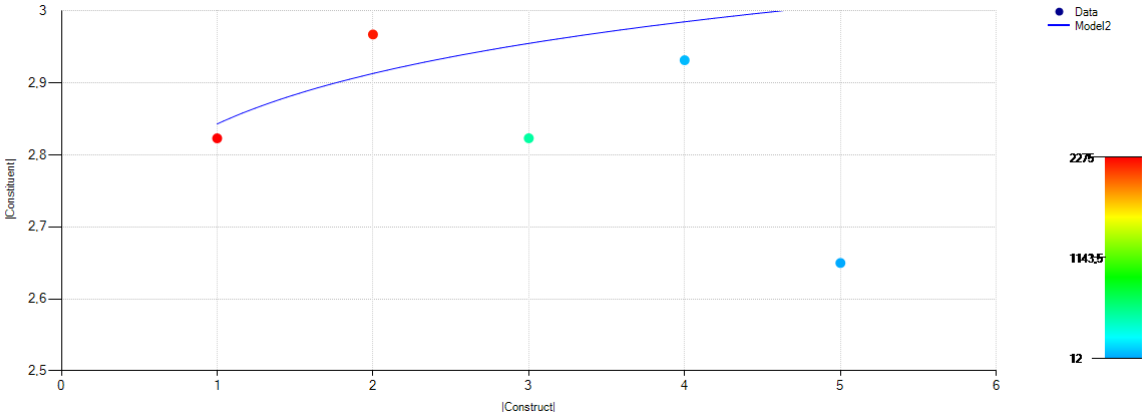


Obrázek 17 Zastoupení *x*-slabičných slov, Výběrové soubory 1–3, alternativní způsob segmentace

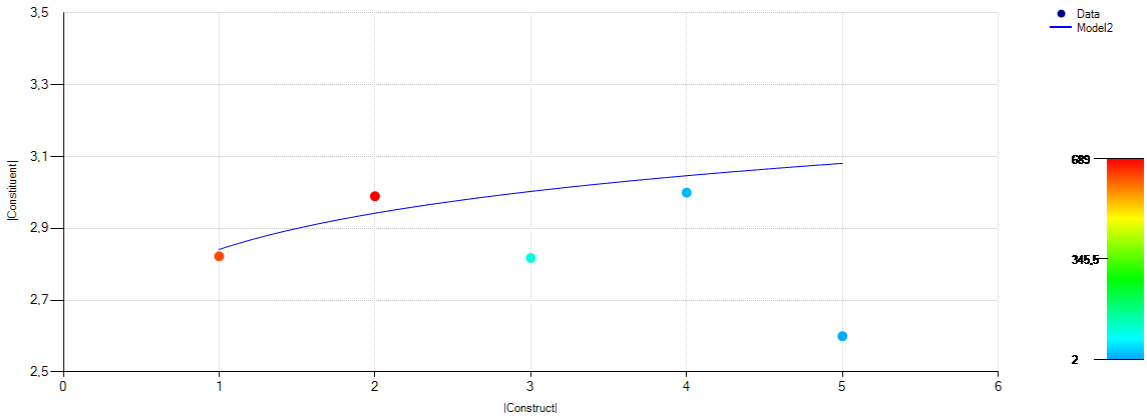
Průměrné hodnoty délek slabik v grafémech se po aplikování alternativního způsobu segmentace pohybují v intervalu ⟨2,60; 3,06⟩, délka intervalu je oproti původní segmentaci kratší, tj. 0,46 (namísto 0,70), což znamená, že délky slov jsou méně rozmanité. U Výběrového souboru 1 lze pozorovat postupný pokles průměrných délek slabik v grafémech (s výjimkou čtyřslabičných slov), avšak u dalších dvou výběrových souborů není klesající tendence patrná, což ilustrují následující grafy.



Obrázek 18 Experiment 1: Grafická vizualizace pozorování uvedených v Tabulce 16 s proloženou regresní křivkou, Výběrový soubor 1: Yu Hua – *Kamarádi* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů) pro námi navržený způsob segmentace



Obrázek 19 Experiment 1: Grafická vizualizace pozorování uvedených v Tabulce 16 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů) pro námi navržený způsob segmentace



Obrázek 20 Experiment 1: Grafická vizualizace pozorování uvedených v Tabulce 16 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů) pro námi navržený způsob segmentace

Oproti původnímu způsobu segmentace se klesající tendence křivky, která znázorňuje nepřímou úměrnost mezi konstrukty a konstituenty definovanou MALem, projevila po segmentování textu alternativním způsobem pouze u Výběrového souboru 1. U ostatních dvou výběrových souborů zůstává křivka rostoucí a parametry *b* nabývají záporných hodnot. V případě Výběrového souboru 1 má koeficient determinace poměrně vysokou hodnotu $R^2 = 0{,}80$, což představuje vysokou shodu. Přesné hodnoty parametrů a koeficientů determinace jsou uvedeny v Tabulce 19. Opět upozorňujeme, že koeficienty determinace R^2 pro Výběrový soubor 2 a 3 se vztahují k stoupající tendenci křivek, což je v rozporu s MALem. Konkrétní hodnoty parametrů *A* i *b* spolu s koeficienty determinace R^2 následují v Tabulce 19.

Tabulka 19 Experiment 1: Parametry *A* a *b* a koeficient determinace R^2 pro matematický model 2 MALu vztahující se k empiricky získaným datům na úrovni U3 pro námi navržený způsob segmentace

Výběrový soubor	**Parametr *A***	**Parametr *b***	***Koeficient determinace* R^2**
Výběrový soubor 1	3,07	0,05	0,80
Výběrový soubor 2	2,84	-0,04	0,34
Výběrový soubor 3	2,84	-0,05	0,46

I když jsou hodnoty parametrů b u Výběrových souborů 2 a 3 záporné, v porovnání s původním způsobem segmentace se hodnoty v obou případech zvýšily směrem k nule. U Výběrového souboru 2 se původní hodnota $b = -0{,}10$ zvýšila na $b = -0{,}04$ a u Výběrového souboru 3 se hodnota $b = -0{,}15$ zvýšila na $b = -0{,}05$. To znamená, že strmější stoupání u původního způsobu segmentace se pomalu dostává k rovnoběžce s osou *x* a kdybychom získali jen o něco lepší data, tendence by se obrátila ke kýženému klesání, jak se stalo v případě Výběrového souboru 1. Třebaže se požadovaná klesající tendence u těchto dvou vzorků neprojevila, rozhodně se zdá, že začleňování vybraných funkčních slov je správným směrem při určování obecné definice čínského slova.

Podívejme se blíže na Výběrový soubor 2 a 3, abychom zjistili, v čem se liší oproti Výběrovému souboru 1.

3.1.2 Experiment 1A

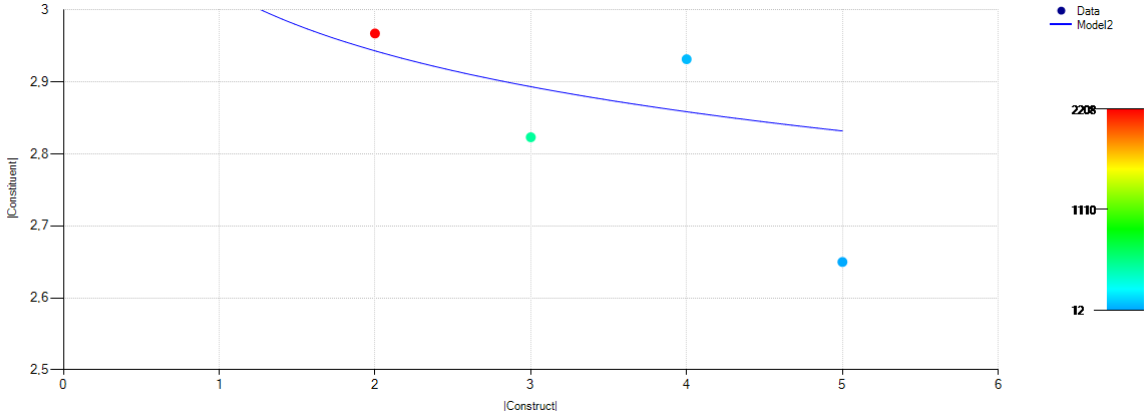
Na první pohled je zřejmé, že klesající tendence u Výběrového souboru 2 a 3 nejvíce narušují opět jednoslabičná slova, která mají vysokou četnost. Podívejme se tedy konkrétněji, která další jednoslabičná slova by mohla způsobovat odchylku.

V případě Výběrového souboru 2 klesající tendenci nejvíce narušuje jednoslabičné slovo *on/ona/ono* tā 他/她/它, které má absolutní četnost výskytu 438 z 2 275 znaků (19,25 %) a je tak nejčetnějším slovem v rámci tohoto výběrového souboru, pokud text segmentujeme alternativním způsobem. Stejně tak v případě Výběrového souboru 3 jsou nečetnější slova osobní zájmena, konkrétně *já* wǒ 我 (79 z 1 432 znaků) a *ty* nǐ 你 (50 z 1 432 znaků). Abychom zjistili, jak moc tato nejčetnější slova ovlivňují výsledky, rozhodli jsme se je v rámci Experimentu 1A vynechat. Toto vynechání provedeme pouze za účelem zjištění, jakou mají váhu v rámci těchto výběrových souborů, a neznamená to, že s nimi již dále nepočítáme. Výsledky získané po odebrání nejčetnějších osobních zájmen uvádíme v Tabulce 20.

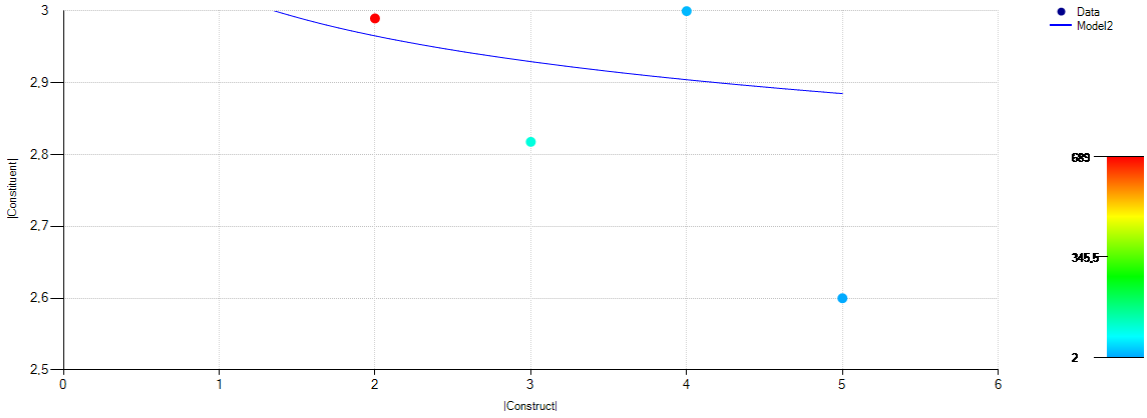
Tabulka 20 Experiment 1A: Jazyková úroveň U3: slovo (měřené v slabikách) – slabika (měřená v průměrném počtu grafémů) pro námi navržený způsob segmentace a po odebrání nejčetnějších osobních zájmen

Výběrový soubor	**Délka slova v slabikách x_3**	**Frekvence z_3**	**Průměrná délka slabik v grafémech y_3**
Výběrový soubor 2	1	1 837	3,02
	2	2 208	2,97
	3	487	2,82
	4	55	2,93
	5	12	2,65
Výběrový soubor 3	1	511	3,02
	2	689	2,99
	3	97	2,82
	4	13	3,00
	5	2	2,60

Z Tabulky 20 je u obou výběrových souborů na první pohled patrný pokles průměrných délek slabik (konstituentů) s rostoucí délkou slov (konstruktů) u nejčetnějších pozorování. Při vynechání nejfrekventovanějších slov by křivka znázorňující vztah mezi konstrukty a jejich konstituenty měla v obou případech klesající tendenci, což potvrzují grafické vizualizace, viz Obrázek 21 a Obrázek 22.



Obrázek 21 Experiment 1A: Grafická vizualizace pozorování uvedených v Tabulce 20 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů) pro námi navržený způsob segmentace a po odebrání osobních zájmen *on/ona/ono* 他/她/它 tā



Obrázek 22 Experiment 1A: Grafická vizualizace pozorování uvedených v Tabulce 20 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů) pro námi navržený způsob segmentace a po odebrání osobních zájmen *já* wǒ 我 a *ty* nǐ 你

Z grafických reprezentací lze pozorovat, že obě křivky mají klesající tendenci a hodnoty parametrů b jsou u obou výběrových souborů kladná čísla. V případě Výběrového souboru 2 hodnota koeficientu determinace R^2 dosahuje relativně vysokých hodnot $R^2 = 0{,}71$, viz Tabulka 21.

Tabulka 21 Experiment 1A: Parametry A a b a koeficient determinace R^2 pro matematický model 2 MALu vztahující se k empiricky získaným datům na úrovni U3 pro námi navržený způsob segmentace a po odebrání nejčetnějších osobních zájmen

Výběrový soubor	**Parametr *A***	**Parametr *b***	***Koeficient determinace R^2***
Výběrový soubor 2	3,03	0,04	0,71
Výběrový soubor 3	3,03	0,03	0,43

U obou zkoumaných výběrových souborů můžeme opět sledovat totožné tendence, avšak u Výběrového souboru 3 je shoda s matematickým modelem MALu výrazně nižší, hodnota koeficientu determinace $R^2 = 0{,}43$. I když je klesající tendence patrná i u tohoto

výběrového souboru, hodnota $R^2 = 0{,}43$ je příliš nízká, abychom ji mohli považovat za potvrzení shody s tendencí MALu.

Jsme si vědomi, že tyto výsledky jsme získali na základě odebrání dat, avšak tento experiment má za cíl zjistit, do jaké míry mohou nejčetnější slova ovlivňovat celkové výsledky. Na základě tohoto experimentu jsme zjistili, že nejčetnější slovo může mít na výsledky zásadní vliv a může ovlivnit chování křivky, respektive tendenci MALu.

Zároveň se ptáme, z jakého důvodu je klesající tendence patrná u Výběrového souboru 1, aniž by bylo nutné odebírat jakákoli data? Pokud se blíže podíváme na jednoslabičná slova, oproti Výběrovému souboru 2 a 3 jsou kromě osobních zájmen hojně používána i vlastní jména dvou hlavních představitelů povídky Kūn Shān a Shí Gāng, a proto výsledky nejsou tolik ovlivněny jednoslabičnými osobní zájmeny *on/ona* tā 他/她, jako tomu bylo například u Výběrového souboru 2. Užívání osobních zájmen *on* nebo *ona* tā 他/她 místo vlastních jmen dokonce obrací tendenci a dělá tak texty neekonomickými. Toto zjištění je velmi překvapivé, protože efektivnější by bylo použít kratší osobní zájmena namísto delších vlastních jmen, aby byla zachována ekonomičnost jazyka (samozřejmě v případě, že je znám kontext).

K zamyšlení zůstává, jaká data bychom získali, kdybychom osobní zájmeno *on/ona* tā 他/她 nahradili vlastním jménem a tím bychom získali více delších jednoslabičných nebo více víceslabičných slov. Předpokládáme, že v tom případě by struktura textu více odpovídala pravidlům, které stanoví MAL. Je to pouze hypotéza, kterou by bylo třeba prověřit.

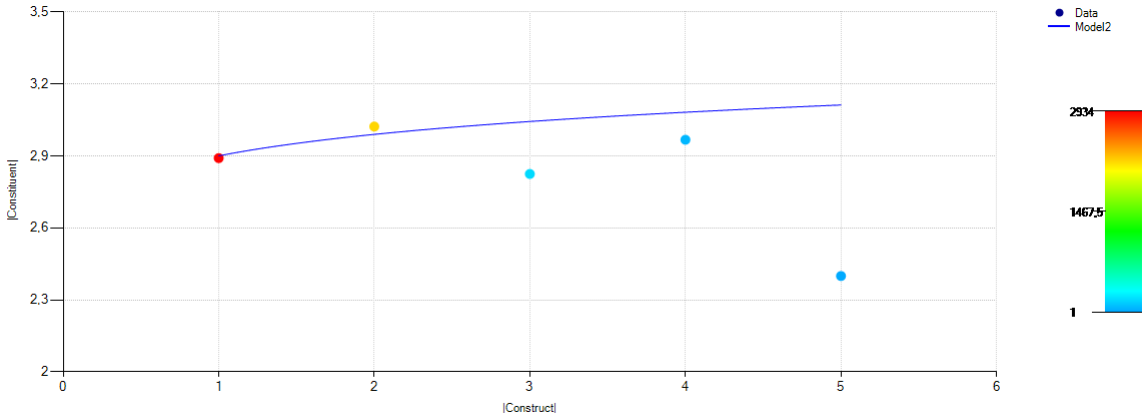
Pro úplnost bude posledním krokem v rámci tohoto experimentu porovnání, jaká by byla shoda s MALem, pokud bychom osobní zájmena odebrali u textů, které jsme segmentovali původním způsobem, viz Tabulka 22.

Tabulka 22 Experiment 1A: Jazyková úroveň U3: slovo (měřené v slabikách) – slabika (měřená v průměrném počtu grafémů) pro **původní způsob** segmentace dle normy GB/T 16159–2012 a po odebrání nejčetnějších osobních zájmen

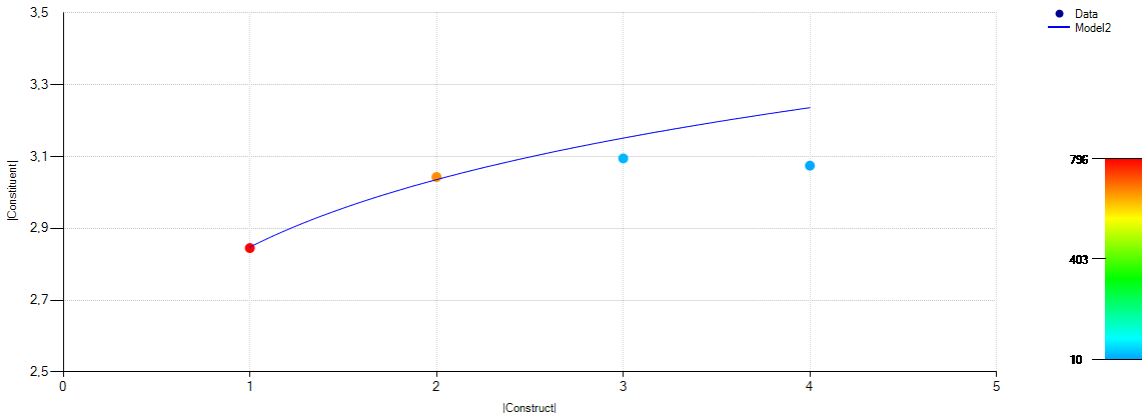
Výběrový soubor	**Délka slova v slabikách x_3**	**Frekvence z_3**	**Průměrná délka slabik v grafémech y_3**
	1	2 934	2,89
	2	2 194	3,02

Výběrový soubor 2	3	166	2,83
	4	47	2,97
	5	1	2,40
Výběrový soubor 3	1	796	2,85
	2	666	3,04
	3	21	3,10
	4	10	3,08

Po odebrání nejčetnějších osobních zájmen u obou zkoumaných výběrových souborů zůstávají nejčetnější jednoslabičná slova. Předpoklad, že s rostoucí délkou konstruktů (slov) klesá průměrná délka konstituentů (slabik), se pro původní způsob segmentace neprojevil ani po odebrání nejčetnějších osobních zájmen, což je na první pohled jasné z dat uvedených v Tabulce 22. Na základě těchto údajů jsme pomocí softwaru MA studio vygenerovali grafické reprezentace, které zobrazují tendenci křivek, viz Obrázek 23 a Obrázek 24.



Obrázek 23 Experiment 1A: Grafická vizualizace pozorování uvedených v Tabulce 22 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů) pro **původní způsob** segmentace dle normy GB/T 16159–2012 a po odebrání osobních zájmen *on/ona/ono* tā 他/她/它



Obrázek 24 Experiment 1A: Grafická vizualizace pozorování uvedených v Tabulce 22 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů) pro **původní způsob** segmentace dle normy GB/T 16159–2012 a po odebrání osobních zájmen *já* wǒ 我 a *ty* nǐ 你

Oproti alternativnímu způsobu segmentace zůstávají tendence křivek stále rostoucí, což je překvapivé, protože u alternativního způsobu segmentace mělo odebrání osobních zájmen zásadní vliv, a dokonce obrátilo rostoucí tendenci na klesající. Jelikož jsou hodnoty parametrů *b* záporné, hodnoty koeficientů determinace nejsou relevantní, ale i tak je pro úplnost připojujeme v Tabulce 23.

Tabulka 23 Experiment 1A: Parametry *A* a *b* a koeficient determinace R^2 pro matematický model 2 MALu vztahující se k empiricky získaným datům na úrovni U3 pro **původní způsob** segmentace dle normy GB/T 16159–2012 a po odebrání nejčetnějších osobních zájmen

Výběrový soubor	**Parametr *A***	**Parametr *b***	**Koeficient determinace R^2**
Výběrový soubor 2	2,90	-0,04	0,52
Výběrový soubor 3	2,85	-0,09	0,98

Co se týče Výběrových souborů 4 a 5, které byly původně zaznamenány latinkou, ani u jednoho z výběrových souborů jsme nezískali data potvrzující shodu s MALem, proto ani ortografická slova v těchto textech nevyhovují ekonomizujícím pravidlům jazyka a na této jazykové hladině neodráží obecně platnou definici čínského slova. Tyto

výběrové soubory jsme již nesegmentovali alternativním způsobem, protože byly zaznamenány latinkou a členění textů tak byla jasně dáno.

Shrnutí

Na jazykové hladině U3 se tendence definované MALem po segmentaci textů dle normy GB/T 16159–2012 neprojevily ani u jednoho výběrového souboru, ať už psaného ve znacích, nebo originálně psaného v pinyinu. Části textu, které jsme získali segmentací dle normy GB/T 16159–2012 a pro účely našeho experimentu jsme je považovali za slova, nesplňují předpoklady MALu, a proto takto definované čínské ortografické slovo v námi zvolených výběrových souborech není v souladu s ekonomizujícími pravidly jazyka a na této jazykové hladině neodráží obecnou definici čínského slova.

Z toho důvodu jsme se rozhodli přistoupit k dalšímu kroku a navrhnuli jsme odlišný způsob segmentace, který připojuje vybraná funkční slova k slovům plnovýznamovým. Tento způsob segmentace vychází jednak z nabízených variantních zápisů, které definuje norma, a také ze způsobu segmentace, kterým byl členěn deníkový záznam psaný v pinyinu (Výběrový soubor 4). Pomocí MALu jsme zjišťovali, jestli námi navržený způsob segmentace přinese lepší výsledky. Na základě pozorovaných dat můžeme konstatovat, že námi navrhovaný způsob segmentace více vyhovuje ekonomizujícím zákonům, což jsme ověřili pomocí MALu. Dále je však nutné tento způsob segmentace aplikovat i na dalších jazykové hladiny, abychom získali kompletní podklady pro vyhodnocení. Mimoto jsme zjistili, že četnost jednoho slova může zásadním způsobem ovlivnit celkové výsledky a může změnit tendenci.

Další možnou příčinou, proč se na této jazykové úrovni neprojevily ekonomizující pravidla jazyka, je, že v abecedě pinyin se vyskytují spřežky a dva grafémy označují jediný zvuk.

Tím, že jsme hypotézu ověřovali pouze na povídkách, je nutné provést více experimentů na různých typech žánrů, abychom ji mohli potvrdit.

3.2 Jazyková úroveň U2: klauze – slovo

V rámci další jazykové úrovně se zaměříme na zkoumání jazykových jednotek klauze a slova. Konstruktem x_2 na této jazykové úrovni je klauze měřená v počtu slov, slovo je jeho konstituentem y_2 a je měřené v průměrném počtu slabik, z_2 představuje

frekvenci klauzí dané délky. Pomocí MALu budeme opět zjišťovat, jestli mezi těmito jednotkami existují ve zvolených výběrových souborech ekonomizující zákonitosti a zda ortografické slovo, které vznikne při převodu čínských znaků do latinky dle normy GB/T 16159–2012, odpovídá obecné definici čínského slova. Stejně jako na předešlé nižší jazykové úrovni budeme i nyní zkoumat všechny Výběrové soubory 1–5.

Výsledky kvantifikace všech výběrových souborů jsou prezentovány v Tabulce 24.

Tabulka 24 Jazyková úroveň U2: klauze (měřená ve slovech) – slovo (měřené v průměrném počtu slabik)

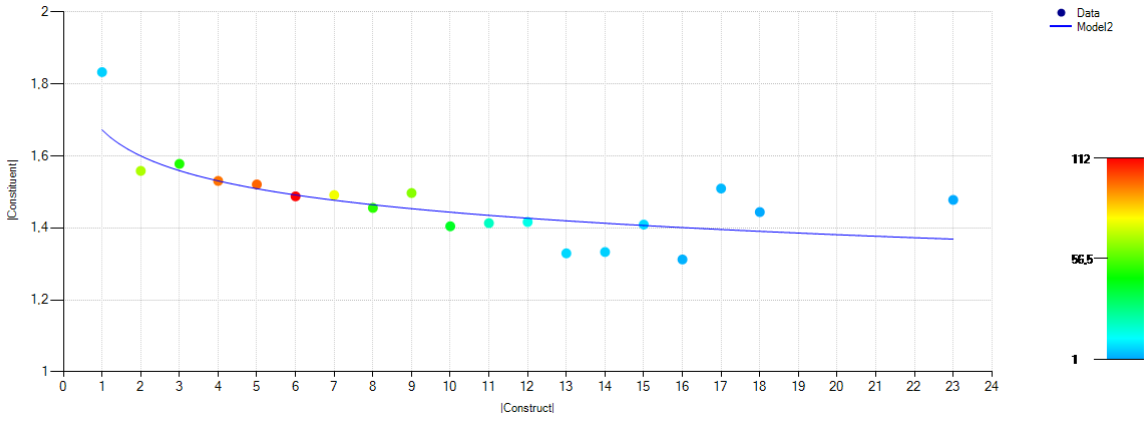
Výběrový soubor	**Délka klauze ve slovech x_2**	**Frekvence z_2**	**Průměrná délka slov v slabikách y_2**
Výběrový soubor 1	1	6	1,83
	2	68	1,56
	3	49	1,58
	4	97	1,53
	5	99	1,52
	6	112	1,49
	7	77	1,49
	8	51	1,46
	9	63	1,50
	10	41	1,40
	11	20	1,41
	12	15	1,42
	13	7	1,33
	14	6	1,33
	15	7	1,41
	16	2	1,31
	17	3	1,51
	18	1	1,44
	23	1	1,48

Výběrový soubor 2	1	26	1,50
	2	76	1,60
	3	122	1,51
	4	147	1,51
	5	140	1,50
	6	135	1,46
	7	110	1,46
	8	90	1,44
	9	58	1,46
	10	59	1,35
	11	21	1,41
	12	11	1,42
	13	5	1,42
	14	2	1,32
	15	2	1,43
	16	2	1,53
	19	1	1,37
Výběrový soubor 3	1	7	1,86
	2	27	1,56
	3	35	1,53
	4	46	1,48
	5	49	1,46
	6	36	1,39
	7	22	1,51
	8	19	1,49
	9	17	1,48
	10	11	1,40
	11	6	1,45
	12	4	1,31
	13	3	1,51
	15	4	1,27

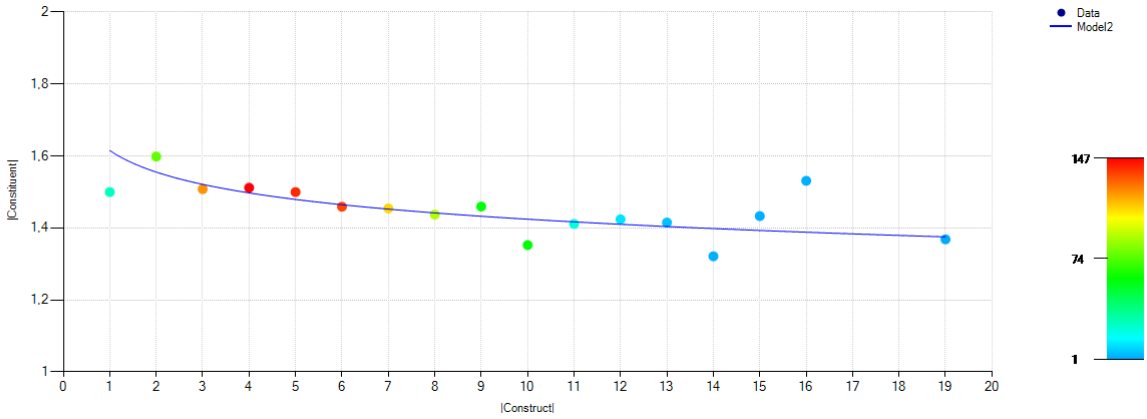
	16	2	1,44
Výběrový soubor 4	1	6	1,83
	2	31	1,79
	3	49	1,67
	4	56	1,58
	5	63	1,57
	6	42	1,55
	7	48	1,53
	8	23	1,51
	9	27	1,49
	10	24	1,48
	11	17	1,46
	12	11	1,52
	13	6	1,49
	14	8	1,49
	15	4	1,57
	16	5	1,40
	17	4	1,44
	20	2	1,43
Výběrový soubor 5	2	4	1,50
	3	10	1,23
	4	16	1,27
	5	9	1,36
	6	12	1,33
	7	9	1,35
	8	5	1,25
	9	7	1,29
	10	4	1,23
	11	1	1,36
	12	1	1,42
	13	1	1,62

Pokud se podíváme na data ze všech Výběrových souborů 1–5, zjistíme, že délka klauzí měřených v počtu slov je od 1 do 23 slov a průměrná délka slov měřených ve slabikách se pohybuje v rozmezí ⟨1,23; 1,86⟩, viz Tabulka 24. S výjimkou Výběrového souboru 2 a 5 je na první pohled patrná klesající tendence průměrných délek konstituentů (slov) se vzrůstající délkou konstruktů (klauzí) u všech výběrových souborů, obzvlášť pokud se zaměříme na nejčetnější pozorování. U méně četných pozorování tato tendence není dodržována a u každého výběrového souboru se setkáváme i s případy, kdy dochází s rostoucí délkou konstruktů také k nárůstu průměrných délek konstituentů, což je v rozporu s tvrzením MALu. Jelikož se však jedná o méně četná pozorování, jsou statisticky méně významná.

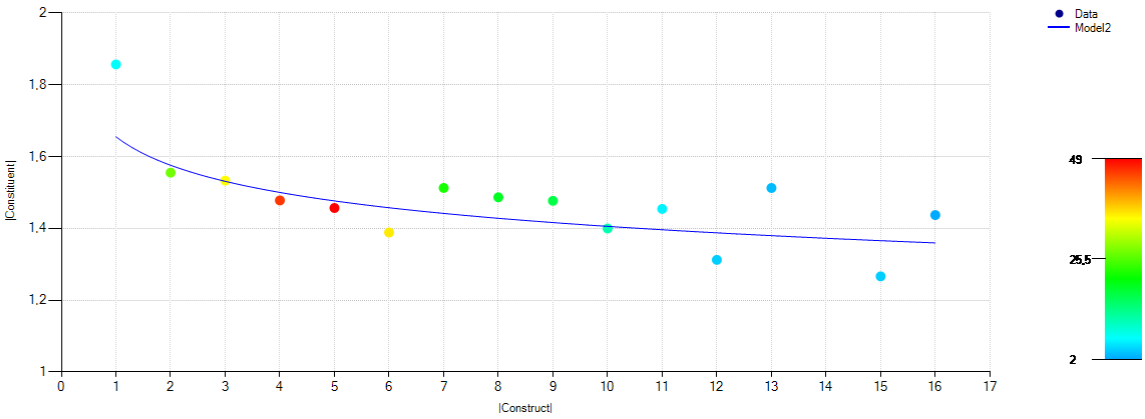
Pro lepší ilustraci získaných dat, uvádíme grafické vizualizace.



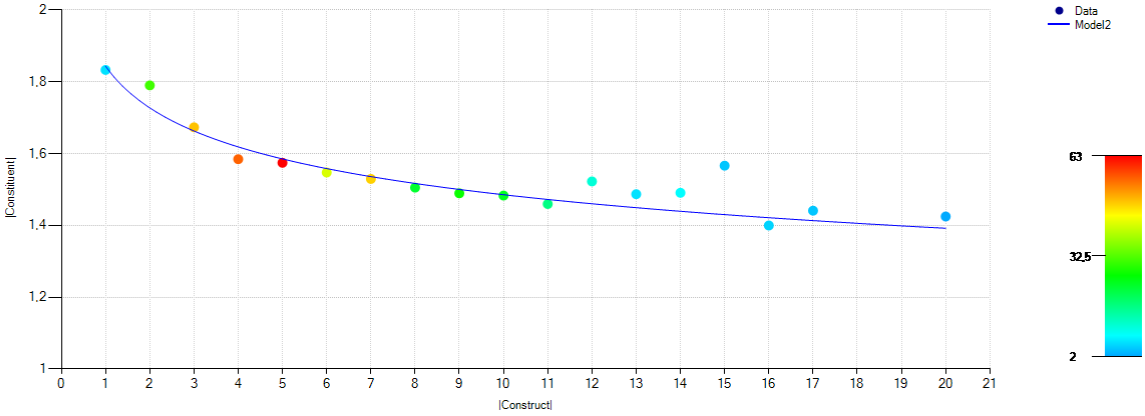
Obrázek 25 Grafická vizualizace pozorování uvedených v Tabulce 24 s proloženou regresní křivkou, Výběrový soubor 1: Yu Hua – *Kamarádi* pro jazykovou úroveň U2 (klauze měřená ve slovech – slovo měřené v průměrném počtu slabik)



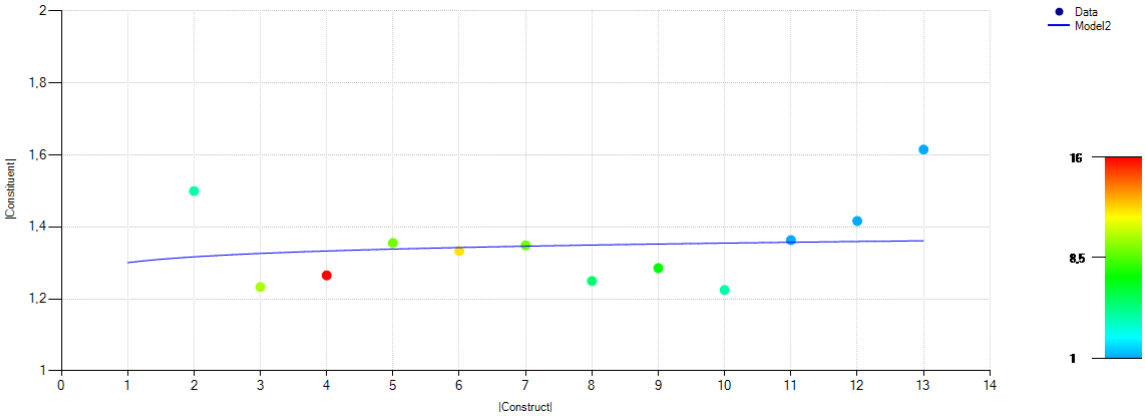
Obrázek 26 Grafická vizualizace pozorování uvedených v Tabulce 24 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U2 (klauze měřená ve slovech – slovo měřené v průměrném počtu slabik)



Obrázek 27 Grafická vizualizace pozorování uvedených v Tabulce 24 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U2 (klauze měřené ve slovech – slovo měřené v průměrném počtu slabik)



Obrázek 28 Grafická vizualizace pozorování uvedených v Tabulce 24 s proloženou regresní křivkou, Výběrový soubor 4: Zhang Liqing – *Deníkové záznamy v pinyinu* pro jazykovou úroveň U2 (klauze měřená ve slovech – slovo měřené v průměrném počtu slabik)



Obrázek 29 Grafická vizualizace pozorování uvedených v Tabulce 24 s proloženou regresní křivkou, Výběrový soubor 5: *Integrated Chinese Level 1 Part 2* pro jazykovou úroveň U2 (klauze měřená ve slovech – slovo měřené v průměrném počtu slabik)

Dle předpokladu mají výsledné tvary křivek kromě Výběrového souboru 5 klesající tendenci a jsou konvexní, tím splňují předpoklad MALu. Stejně tak hodnoty parametrů *b* pro všechny výběrové soubory s výjimkou Výběrového souboru 5 nabývají kladných hodnot. Výběrový soubor 5 má však v porovnání s ostatními výběrovými soubory mnohem nižší počet konstruktů, tj. klauzí (viz porovnání v Tabulce 25), proto je možné, že tento výběrový soubor nedisponuje dostatečným množstvím dat, aby se vztahy mezi konstrukty a konstituenty mohly projevit, a statistická relevantnost tak není vysoká. Hodnoty parametrů *A* a *b* a koeficientů determinace R^2 jsou uvedeny v Tabulce 26. Normalita a homoskedasticita jsou splněny.

Tabulka 25 Celkový počet klauzí, Výběrové soubory 1–5

Výběrový soubor	Celkový počet klauzí
Výběrový soubor 1	725
Výběrový soubor 2	1 007
Výběrový soubor 3	288
Výběrový soubor 4	426
Výběrový soubor 5	79

Tabulka 26 Parametry *A* a *b* a koeficient determinace R^2 pro matematický model 2 MALu vztahující se k empiricky získaným datům na úrovni U2

Výběrový soubor	Parametr *A*	Parametr *b*	*Koeficient determinace* R^2
Výběrový soubor 1	1,67	0,06	0,74
Výběrový soubor 2	1,62	0,06	0,65
Výběrový soubor 3	1,66	0,07	0,53
Výběrový soubor 4	1,84	0,09	0,89
Výběrový soubor 5	1,30	-0,02	0,02

Koeficienty determinace R^2 mají poměrně vysokou hodnotu u Výběrového souboru 1 ($R^2 = 0{,}74$) a Výběrového souboru 4 ($R^2 = 0{,}89$), což značí shodu s matematickým modelem MALu. Také hodnoty $R^2 = 0{,}65$ a $R^2 = 0{,}53$ naznačují, že i

ve Výběrovém souboru 2 a 3 se v určité míře projevuje tendence definovaná MALem, avšak hodnota není dostačující pro potvrzení shody. Nejprve se podívejme, z jakého důvodu se projevila nižší shoda s tendencí MALu u Výběrového souboru 2.

U Výběrového souboru 2 klesající tendenci viditelně narušuje hned první pozorování, tedy klauze o délce jednoho slova. Ostatní pozorování, která výrazně narušují klesající tendenci, mají nižší četnost a neovlivňují tolik výsledek, proto se jimi nebudeme zabývat. Při bližším zkoumání jednoslovných klauzí jsme zjistili, že z 26 klauzí se 15 z nich skládá z jednoslabičného slova, 10 z nich z dvouslabičných slov a pouze jedna klauze obsahuje čtyřslabičné slovo. Právě jednoslabičná slova, která tvoří jednoslovnou klauzi, odporují předpokladu, že čím je délka klauze měřená ve slovech kratší, tím je délka slov měřených v průměrném počtu slabik delší. Jednoslovné klauze by se tedy měly skládat z víceslabičných slov. Podíváme-li se na konkrétní případy jednoslovných klauzí, zjistíme zajímavý fakt, že ve všech případech se jedná o přímou řeč a nejfrekventovanějším výskytem je slovo *haló* wèi 喂, které bylo použito celkem 8x (z 15 klauzí). Dále je následováno záporkou bù 不 (výskyt celkem 4x) a osobním zájmenem *ty* nǐ 你, které se vyskytlo celkem 2x a figuruje pouze v nedokončených výpovědích. Níže uvádíme konkrétní příklady uvedených slov (zvýrazněné tučně):

Haló, haló, slyšíš? **Wèi, wèi,** nǐ tīngdào le ma? **喂，喂**，你听到了吗？

Haló. **Wèi. 喂。**

Jsi nestydatý, jsi odporný, jsi hnusný, jsi … Nǐ wúchǐ, nǐ bēibǐ, nǐ xiàliú, **nǐ**…… 你无耻，你卑鄙，你下流，**你……**

Haló, haló, kdo je to? Proč není slyšet zvuk…? **Wèi, wèi,** shì shéi? Zěnme méiyǒu shēngyīn…… **喂，喂**，是谁？怎么没有声音……

Ó, to je Lin Hong… **Ō,** shì Lín Hóng…… **噢**，是林红……

Ty… **Nǐ**…… **你……**

Ne! **Bù**！**不！**

Ne. **Bù. 不。**

Ne. **Bù. 不。**

Ne, **Bù, 不，**

Pravděpodobně se jedná o specifikum tohoto výběrového souboru, a proto je možné, že právě přímé řeči, zejména pokud se jedná o nedokončené výpovědi, narušují ekonomičnost jazyka.

U Výběrového souboru 3 se klesající tendence zpočátku projevila, ale v bodě $x_2 = 7$, tj. klauze sestávající z 7 slov a více, je klesající tendence narušena. Rostoucí tendence je však pouze lokální (k nárůstu došlo pouze v úseku $x_2 = 6$ a $x_2 = 7$), poté se zase projevuje klesající tendence. Následující body, které jsou roztroušeny poměrně nahodile, mají nízkou frekvenci, a proto nejsou statisticky významné.

Nejvyšší shodu s matematickým modelem MALu je možné pozorovat u Výběrového souboru 4, který byl původně zaznamenán v pinyinu čínskou autorkou a u kterého jsme grafickou podobu slova nijak neupravovali. Důvodem, proč toto členění textu na slova lépe odpovídá předpokladům MALu, může být, že délka slov zaznamenaných v pinyinu je viditelná na první pohled a autorka tomu přizpůsobuje i délky klauzí a souvětí. Naopak při psaní ve znacích autoři nezohledňují tolik zvukovou stránku, kterou určitým způsobem odráží pinyin, a délka vyslovených slov není na první pohled viditelná. Další možností je, že konkrétně tato autorka při psaní sleduje tendence definované MALem, a proto je třeba tuto domněnku ověřit na více výběrových souborech psaných v pinyinu. Důvodem, proč se podobná tendence neprojevila i u Výběrového souboru 5, který byl také zaznamenán latinkou, může být nedostatečná délka výběrového souboru, ale i odlišný způsob segmentování textu na slova.

Na jazykové úrovni U2, kde slovo představuje konstituent klauzí, jsme získali rozporuplné výsledky. Dle statistické metody jsme pouze u dvou výběrových souborů (Výběrový soubor 1 a 4) ověřili, že ortografické slovo (vymezené dle normy GB/T 16159–2012) vyhovuje ekonomickým pravidlům jazyka. U dalších dvou výběrových souborů (Výběrový soubor 2 a 3) byla sice požadovaná klesající tendence naznačena, avšak hodnota koeficientu determinace byla nižší než 0,70, a nelze tak uvažovat o adekvátním a dobře sedícím modelu. U Výběrového souboru 5 (učebnicové texty) se shoda neprokázala vůbec. Proto nelze prohlásit, že čínské ortografické slovo v případě těchto výběrových souborů sleduje ekonomická pravidla jazyka a je v souladu s obecně platnou definicí slova.

Na předešlé jazykové úrovni U3 jsme navrhli alternativní způsob segmentace slov a abychom mohli ověřit, jestli tento způsob členění textu zaznamenaného v pinyinu více vyhovuje ekonomizujícím zákonům v jazyce, je žádoucí ho aplikovat i na další jazykovou

úroveň. Tentokrát bude slovo stát na pozici konstituentu a s rostoucí délkou klauzí (měřených ve slovech) by měly klesat průměrné délky slov (měřených ve slabikách). Segmentace klauzí musí zůstat stále stejná.

3.2.1. Experiment 2

V experimentu 2 budeme opět zkoumat první tři výběrové soubory, které byli původně psané v čínských znacích, a převedeme je do pinyinu dle námi navrženého způsobu segmentace (blíže viz 3.1.1 Experiment 1). Výsledky nám napoví, který způsob segmentace na slova více vyhovuje ekonomizujícím pravidlům. Data získaná po aplikaci alternativního způsobu segmentace jsou uvedena v následující Tabulce 27.

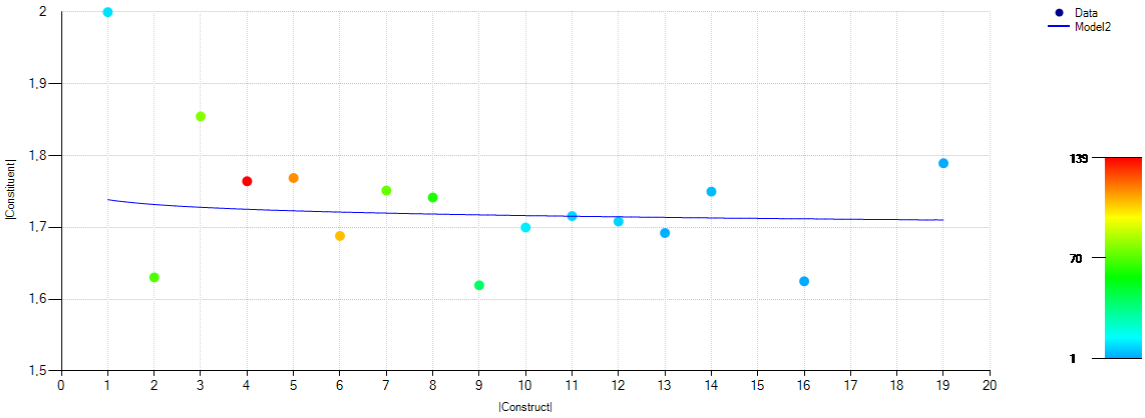
Tabulka 27 Experiment 2: Jazyková úroveň U2: klauze (měřená ve slovech) – slovo (měřené v průměrném počtu slabik) pro námi navržený způsob segmentace

Výběrový soubor	**Délka klauze ve slovech x_2**	**Frekvence z_2**	**Průměrná délka slov v slabikách y_2**
Výběrový soubor 1	1	10	2,00
	2	69	1,63
	3	77	1,85
	4	139	1,76
	5	116	1,77
	6	108	1,69
	7	72	1,75
	8	60	1,75
	9	40	1,62
	10	11	1,70
	11	8	1,72
	12	8	1,71
	13	2	1,69
	14	3	1,71
	16	1	1,63
	19	1	1,79
	1	26	1,54

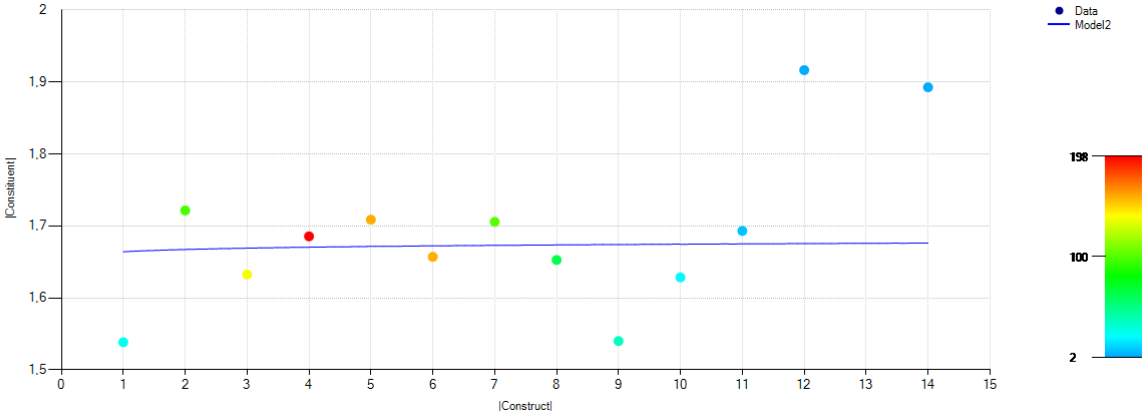
Výběrový soubor 2	2	97	1,72
	3	137	1,63
	4	198	1,69
	5	158	1,71
	6	158	1,66
	7	101	1,71
	8	63	1,65
	9	36	1,54
	10	21	1,63
	11	8	1,69
	12	2	1,92
	14	2	1,89
Výběrový soubor 3	1	8	2,00
	2	31	1,63
	3	43	1,64
	4	61	1,63
	5	47	1,63
	6	27	1,56
	7	25	1,73
	8	24	1,65
	9	10	1,81
	10	4	1,58
	11	3	1,61
	12	2	1,63
	13	1	1,54
	14	2	1,64

Při alternativním způsobu segmentace se u všech výběrových souborů snížily délky klauzí ve slovech, tj. od 1 do 19 (místo původní nejdelší klauze obsahující 23 slov). Získali jsme tedy delší slova, jejichž průměrná délka měřená ve slabikách se nově pohybuje v rozmezí $\langle 1{,}54; 2{,}00\rangle$ (původní rozmezí bylo $\langle 1{,}23; 1{,}86\rangle$). Z Tabulky 27 není na první pohled ani u jednoho výběrového souboru patrné, že by s rostoucí délkou

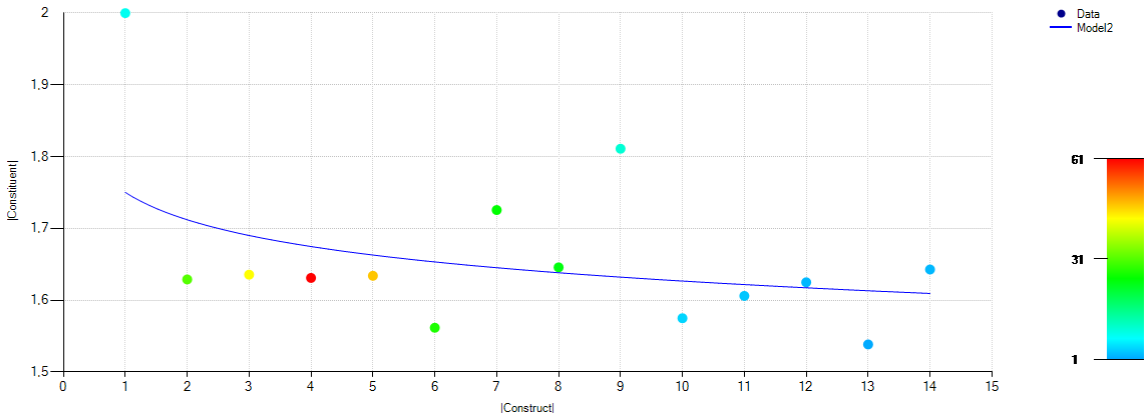
klauzí (měřených ve slovech) klesala průměrná délka slov (měřených ve slabikách). Proto se podívejme na grafické vizualizace, které lépe naznačí, zda získaná data vyhovují předpokladu MALu a zda slovo definované tímto způsobem více vyhovuje ekonomizujícím pravidlům jazyka, viz Obrázek 30, Obrázek 31 a Obrázek 32.



Obrázek 30 Experiment 2: Grafická vizualizace pozorování uvedených v Tabulce 27 s proloženou regresní křivkou, Výběrový soubor 1: Yu Hua – *Kamarádi* pro jazykovou úroveň U2 (klauze – slovo) pro námi navržený způsob segmentace



Obrázek 31 Experiment 2: Grafická vizualizace pozorování uvedených v Tabulce 27 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U2 (klauze – slovo) pro námi navržený způsob segmentace



Obrázek 32 Experiment 2: Grafická vizualizace pozorování uvedených v Tabulce 27 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U2 (klauze – slovo) pro námi navržený způsob segmentace

Z uvedených grafických vizualizací lze vypozorovat, že jednotlivé body jsou rozprostřeny poměrně nahodile a klesající tendence není patrná – snad pouze s výjimkou Výběrového souboru 3, u kterého je evidentní alespoň náznak požadované tendence. Ačkoli parametry b mají kladné hodnoty, hodnoty koeficientu determinace R^2 jsou ve všech případech minimální a v porovnání s původním způsobem segmentace došlo k velkému snížení hodnot. U Výběrového souboru jsme z původní hodnoty $R^2 = 0{,}74$ získali mnohem nižší hodnotu $R^2 = 0{,}01$, u Výběrového souboru 2 byla původní hodnota $R^2 = 0{,}65$ a po aplikování alternativního způsobu segmentace $R^2 = 0{,}14$, u Výběrového souboru 3 byla původní hodnota $R^2 = 0{,}53$ a nově se snížila na $R^2 = 0{,}11$, viz Tabulka 26 a Tabulka 28.

Tabulka 28 Experiment 2: Parametry *A* a *b* a koeficient determinace R^2 pro matematický model vztahující se k empiricky získaným datům na úrovni U2 pro námi navržený způsob segmentace

Výběrový soubor	**Parametr *A***	**Parametr *b***	***Koeficient determinace R^2***
Výběrový soubor 1	1,74	0,01	0,01
Výběrový soubor 2	1,66	0,00	0,14
Výběrový soubor 3	1,75	0,03	0,11

U Výběrového souboru 1 je nejpatrnější narušení tendence u $x_2 = 2$, tedy u klauzí sestávající ze dvou slov. U těchto klauzí mají slova průměrně kratší délku měřenou

ve slabikách – ideálně by se měla pohybovat mezi 2,00 a 1,85, ale je pouhých 1,63. Tento průměr ovlivňují klauze složené ze dvou jednoslabičných slov (jejich průměrná délka ve slabikách je 1,00) nebo z jednoho jednoslabičného a jednoho dvouslabičného slova (jejich průměrná délka je 1,50). Ze všech klauzí složených ze dvou jednoslabičných slov se nejčastěji opakuje věta: *On řekl.* Tā shuō. 他说。, která se z celkových 10 případů objevila 5x. Klauze složené z jednoho jednoslabičného a jednoho dvouslabičného slova byla ve skupině dvouslovných klauzí nejčetnější, celkem se vyskytlo 64 takto tvořených klauzí. Zde se nejčastěji opakovala spojení: *Já jsem viděl.* Wǒ kàndao. 我看到。(10x), *Myslel jsem si.* Wǒ xīnxiǎng. 我心想。(6x), *Uslyšel jsem.* Wǒ tīngdào. 我听到。(4x) apod.

Při bližším zkoumání těchto klauzí jsme však zjistili, že alternativní způsob segmentace téměř všech klauzí složených ze dvou jednoslabičných slov nebo z jednoho jednoslabičného a jednoho dvouslabičného slova se shoduje s původním způsobem segmentace, proto bychom měli získat podobné výsledky. Stejně jako u původního způsobu segmentace je patrný odskok, avšak u tohoto způsobu segmentace je ještě mnohem viditelnější.

Z toho vyvozujeme, že problémovým pozorováním musí být klauze o délce 3 slov, u kterých se po aplikování alternativního způsobu segmentace výrazně navýšily průměrné hodnoty délek slov ve slabikách (z hodnoty 1,58 na 1,85). Toto navýšení délek slov je způsobené právě připojováním funkčních slov k plnovýznamovým slovům a týká se mnoha případů, proto je zde nelze všechny konkretizovat. Jedná se například o připojování záporky bù 不, slovce de 的, záložek shàng 上 a lǐ 里.

Důvodem, proč se u Výběrového souboru 1 u alternativního způsobu segmentace neprojevila tendence, kterou předpokládá MAL, může být spojení dvou výše uvedených důvodů, které vychází ze stavby tohoto konkrétního textu. Za prvé to mohou být nízké průměrné délky slov u dvouslovných klauzí způsobené častým použitím krátkých slov, což narušuje výsledky nejenom u alternativního způsobu segmentace, ale také u původního způsobu segmentace, i když v menší míře. Za druhé alternativní způsob segmentace výrazně navýšil průměrné délky slov u tříslovných klauzí, proto je možné, že tento způsob segmentace neodpovídá předpokladům MALu.

Výběrovému souboru 2 jsme se věnovali již u původního způsobu segmentace, kdy jsme si všimli, že nižší shoda s tendencí MALu byla pravděpodobně způsobena použitím přímých řečí v souvislosti s nedokončenými výpověďmi. Toto specifikum

ovlivňuje výsledky i při alternativním způsobu segmentace a shoda s tendencí MALu se dokonce ještě snížila, což může být samozřejmě podpořeno i nevhodným způsobem segmentace.

Také u Výběrového souboru 3 jsme po segmentování textu alternativním způsobem obdrželi výsledky, které neprokázaly shodu s MALem. Na základě získaných dat můžeme prohlásit, že alternativní způsob segmentace čínského textu na slova na jazykové úrovni U2 klauze – slovo nepřinesl lepší výsledky ani pro jeden výběrový soubor, který byl předmětem zkoumání. Důvodem může být zvolení nevhodného způsobu segmentace slov, případně nevyhovující členění souvětí na klauze. Je nutné zvážit, jestli by definice klauze neměla být založena výhradně na použití interpunkčních znamének. V našem výzkumu sice interpunkce sloužila jako pomocný indikátor k segmentaci klauzí, ale hranice klauzí netvořila ve všech případech. Proto v dalším výzkumu navrhujeme klauzi definovat pomocí interpunkčních znamének, kdy hranicí klauze v souvětí bude čárka nebo středník. Tento způsob definice klauze používají i čínští vědci, viz například studie Hou et al. (2017) a Chen a Liu (2022).

Shrnutí

Po vyhodnocení dat získaných na jazykové úrovni U2 jsme zjistili, že pouze výsledky ze dvou výběrových souborů (Výběrový soubor 1 a 4) potvrzují hypotézu, že ortografická slova, které vzniknou po segmentaci textu dle normy GB/T 16159–2012, jsou v souladu s ekonomizujícími zákony a odráží tak obecnou definici čínského slova. V těchto případech se potvrdilo, že s rostoucí délkou konstruktů (klauzí), klesá průměrná délka konstituentů (slov). U dalších dvou výběrových souborů byla sice klesající tendence naznačena, avšak shoda s matematickým modelem MALu, pomocí kterého jsme ověřovali platnost naší hypotézy, byla příliš nízká. V případě učebnicového textu (Výběrový soubor 5) dokonce žádná. Přesto tento způsob segmentace nezavrhujeme, avšak je nutné provézt více experimentů. Je možné, že zvolené výběrové soubory mají specifickou stavbu a výsledky jsou ovlivněny například přímými řečmi s nedokončenou výpovědí.

Nejlepší výsledky jsme získali u Výběrového souboru 4, tedy u deníkového záznamu originálně psaného pinyinem, z čehož usuzujeme, že důležitou roli hraje i grafika textů psaných v pinyinu (viz výše).

Dále jsme na této úrovni ověřovali, jestli začleňování vybraných funkčních slov k plnovýznamovým slovům přinese lepší výsledky. Ani u jednoho ze třech zkoumaných

výběrových souborů jsme nezískali data, která by tento předpoklad potvrdila. Ve všech třech případech se shoda s MALem snížila, což může být opět způsobeno volbou výběrových souborů, které obsahují mnoho přímých řečí s nedokončenými výpověďmi. Právě ony by mohly narušovat ekonomičnost jazyka. Dalším možným vysvětlením je, že alternativní způsob segmentace textu nepředstavuje vhodnější způsob členění textu na slova, což je však třeba nadále ověřit.

V dalším výzkumu je třeba dále zkoumat oba způsoby segmentace na větším vzorku výběrových souborů. Jelikož časté používání přímých řečí může výrazným způsobem ovlivňovat výsledky, v příštím výzkumu se zaměříme na výběrové soubory jiného žánrového zaměření.

Dále navrhujeme zaměřit se na analýzu více textů psaných originálně v pinyinu a zjistit, jestli členění textů na slova obecně odpovídá ekonomickým pravidlům jazyka, stejně jako se ukázalo v našem případě.

Další možnou cestou výzkumu je zvolit odlišnou definici klauze stejně jako ji používají i jiní čínští vědci, viz například studie Hou et al. (2017) a Chen a Liu (2022). Autoři klauzi definují podle počtu čárek a středníků v souvětí, což jistě stojí za ověření.

3.3 Jazyková úroveň U1: souvětí – klauze

Na nejvyšší jazykové úrovni U1 budeme zkoumat vztah mezi jazykovými jednotkami souvětí a klauze. Slovo na této jazykové úrovni není konstruktem ani konstituentem, ale slouží pouze k výpočtu průměrných délek klauzí. Proto tato kapitola není pro předkládanou práci stěžejní a nejvyšší jazykové úrovni U1 nebudeme věnovat tolik pozornosti jako předchozím dvěma nižším jazykovým úrovním. Konstruktem x_1 je na této úrovni souvětí, které je měřené v počtu klauzí, z_1 je frekvence souvětí dané délky a konstituentem y_1 je klauze měřená v průměrném počtu bezprostředně nižší jazykové jednotky, tedy v průměrném počtu slov.

Předmětem segmentace jsou všechny výběrové soubory 1–5 a výsledky jejich kvantifikace jsou uvedeny v Tabulce 29.

Tabulka 29 Jazyková úroveň U1: souvětí (měřené v klauzích) – klauze (měřená v průměrném počtu slov)

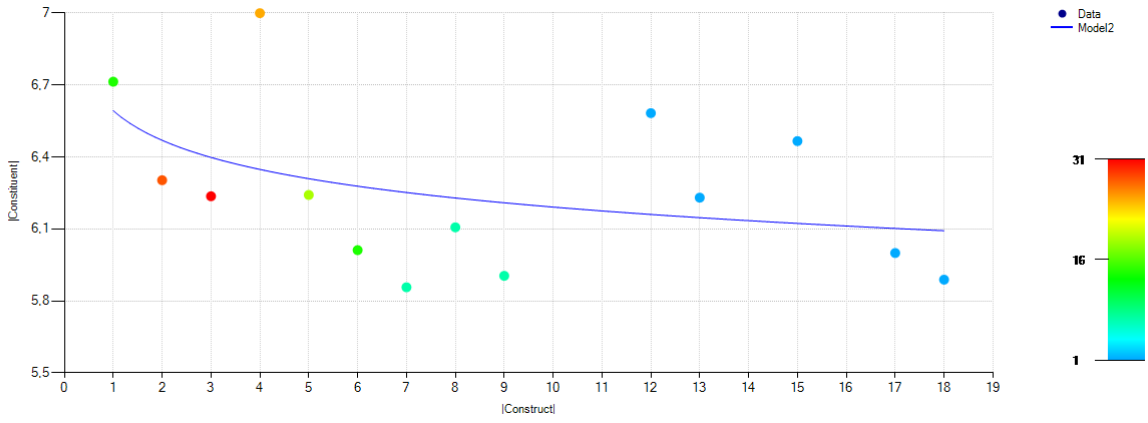
Výběrový soubor	**Délka souvětí v klauzích x_1**	**Frekvence z_1**	**Průměrná délka klauze ve slovech y_1**
Výběrový soubor 1	1	14	6,71
	2	28	6,30
	3	31	6,24
	4	25	7,00
	5	19	6,24
	6	14	6,01
	7	7	5,86
	8	7	6,11
	9	7	5,90
	10	4	7,03
	12	1	6,58
	13	1	6,23
	15	1	6,47
	17	1	6,00
	18	1	5,89
Výběrový soubor 2	1	28	4,18
	2	32	5,53
	3	27	5,06
	4	30	5,41
	5	23	6,50
	6	16	5,47
	7	16	6,17
	8	8	5,44
	9	4	6,53
	10	6	6,15
	11	4	5,73

	12	4	5,29
	13	1	6,31
	14	5	5,81
	16	1	5,75
	17	1	7,00
	23	1	5,65
Výběrový soubor 3	1	10	6,70
	2	8	5,50
	3	12	5,75
	4	9	5,17
	5	8	5,35
	6	7	5,17
	7	2	5,43
	8	2	5,69
	9	2	6,00
	10	2	7,40
	11	1	6,36
	13	1	6,15
	16	1	4,56
Výběrový soubor 4	1	70	8,91
	2	58	6,92
	3	47	5,74
	4	21	5,31
	5	3	5,60
Výběrový soubor 5	1	2	9,50
	2	5	6,80
	3	8	6,17
	4	6	5,13
	5	1	7,00
	6	1	6,33
	8	1	3,75

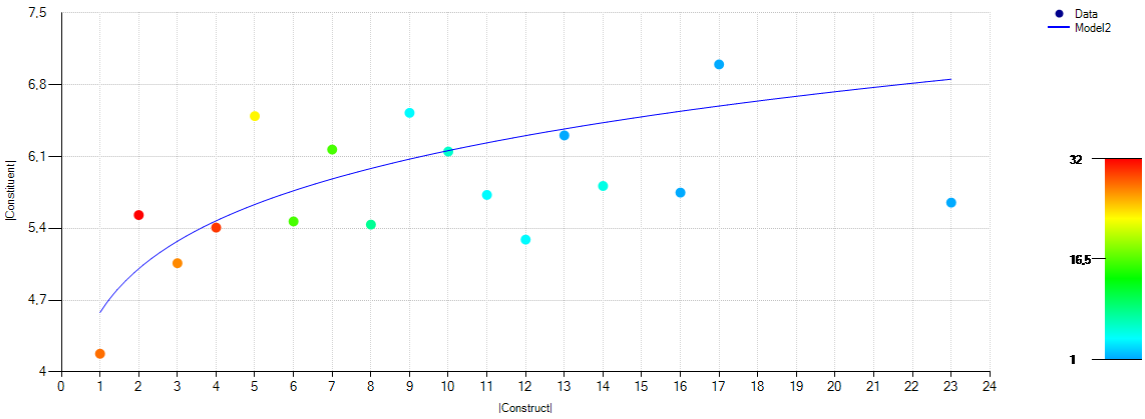
Data z Tabulky 29 ukazují, že se souvětí skládají od 1 do 23 klauzí a průměrná délka klauzí měřených ve slovech se pohybuje v intervalu $\langle 3{,}75; 9{,}50 \rangle$. Na vyšších jazykových úrovních se obecně můžeme potýkat s problémem nedostatečného množství dat, protože se jedná o konstrukty s větší délkou a těch je logicky ve výběrových souborech méně. Tento jev se projevil i na této jazykové úrovni a frekvence nejčetnějšího konstruktu je $z_1 = 70$, což je oproti nižším jazykovým úrovním, kde jsme například na nejnižší úrovni U3 měli frekvenci nejčetnějšího konstruktu slovo $z_3 = 3\,446$, poměrně málo. Podotýkáme, že četnosti mohou ovlivňovat relevantnost výsledků, protože námi zvolený model MALu bere v potaz právě četnosti, kterých je na této úrovni velmi málo. I přes nedostatečné množství vstupních dat u některých výběrových souborů se na tuto jazykovou úroveň blíže podíváme, ale musíme mít na paměti, že tendence nemusí mít stoprocentně vypovídací hodnotu.

Klesající tendenci průměrných délek klauzí s rostoucí délkou souvětí můžeme pozorovat pouze u Výběrového souboru 4 a 5. U ostatních výběrových souborů není na první pohled patrná. I když je klesající tendence narušena i u Výběrových souborů 4 a 5, jedná se vždy o pozorování s nižším počtem výskytů, proto jsou statisticky méně významná.

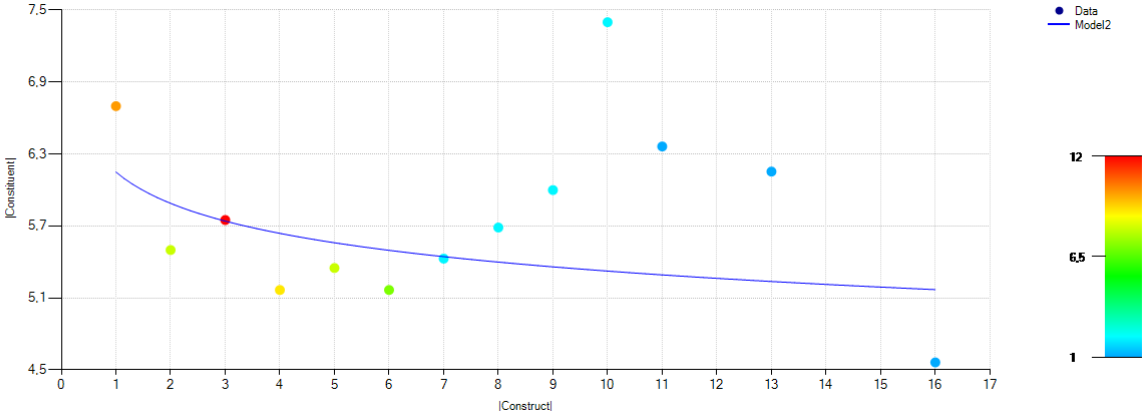
Následné grafické vizualizace poskytnou lepší představu o vztazích mezi konstrukty a jejich konstituenty, viz Obrázek 33 – Obrázek 37.



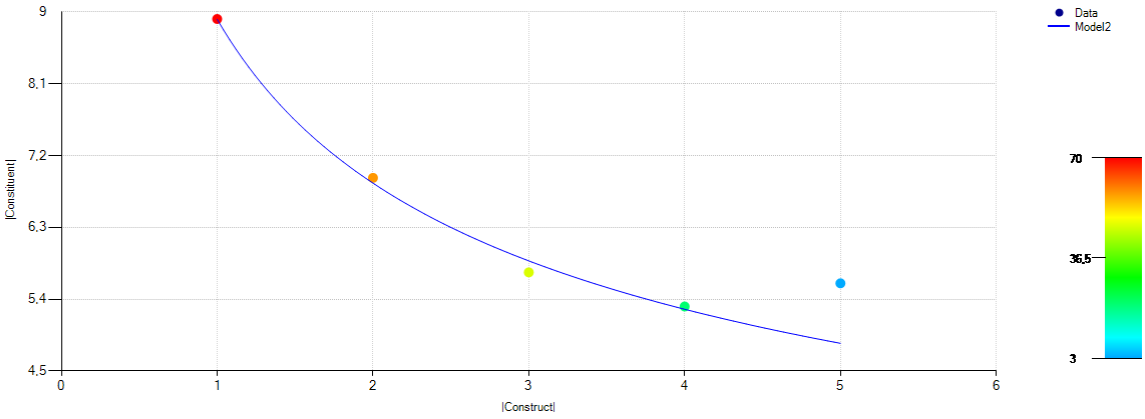
Obrázek 33 Grafická vizualizace pozorování uvedených v Tabulce 29 s proloženou regresní křivkou, Výběrový soubor 1: Yu Hua – *Kamarádi* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřená v průměrném počtu slov)



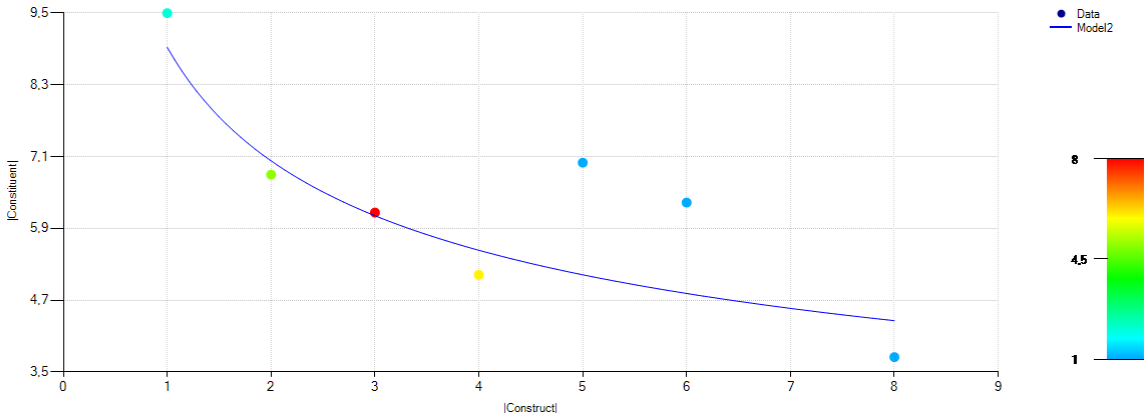
Obrázek 34 Grafická vizualizace pozorování uvedených v Tabulce 29 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřená v průměrném počtu slov)



Obrázek 35 Grafická vizualizace pozorování uvedených v Tabulce 29 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřená v průměrném počtu slov)



Obrázek 36 Grafická vizualizace pozorování uvedených v Tabulce 29 s proloženou regresní křivkou, Výběrový soubor 4: Zhang Liqing – *Deníkové záznamy v pinyinu* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřená v průměrném počtu slov)



Obrázek 37 Grafická vizualizace pozorování uvedených v Tabulce 29 s proloženou regresní křivkou, Výběrový soubor 5: *Integrated Chinese Level 1 Part 2* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřené v průměrném počtu slov)

Z grafických vizualizací je patrné, že kromě Výběrového souboru 2 mají všechny křivky klesající tendenci a splňují předpoklad MALu. Avšak hodnoty koeficientů determinace R^2 jsou pro Výběrový soubor 1 a 3 jsou velmi nízké a opět upozorňujeme, že v případě Výběrového souboru 2 je hodnota platná pro stoupající tendenci křivky, což je v rozporu s tvrzením MALu. U Výběrových souborů 4 a 5 lze pozorovat poměrně vysokou hodnotu koeficientu determinace R^2. Konkrétní hodnoty všech koeficientů determinace R^2 a parametrů *A* a *b* jsou uvedeny v následující Tabulce 30. Normalita a homoskedasticita jsou splněny.

Tabulka 30 Parametry *A* a *b* a koeficient determinace R^2 pro matematický model 2 MALu vztahující se k empiricky získaným datům na úrovni U1

Výběrový soubor	**Parametr *A***	**Parametr *b***	***Koeficient determinace R^2***
Výběrový soubor 1	6,59	0,03	0,10
Výběrový soubor 2	4,58	-0,13	0,56
Výběrový soubor 3	6,15	0,06	0,19
Výběrový soubor 4	8,91	0,38	0,99
Výběrový soubor 5	8,93	0,34	0,73

Křivka, která zachycuje vztah mezi konstrukty a jejich konstituenty a nejlépe odpovídá předpokladům MALu, je nejpatrnější u Výběrového souboru 4 a poměrně přesně kopíruje ideální křivku. U tohoto výběrového souboru dosahuje hodnota koeficientu determinace dokonce hodnoty $R^2 = 0{,}99$, což značí extrémní shodu s matematickým modelem MALu. Jedná se o text, který byl zaznamenán čínskou autorkou v podobě latinky a délka výběrového souboru je poměrně dostačující, abychom mohli prohlásit, že tendence definovaná MALem se u tohoto výběrového souboru projevila. Taktéž Výběrový soubor 5, který byl zaznamenán latinkou, vykazuje klesající tendenci a shodu s MALem ($R^2 = 0{,}73$). U tohoto výběrového souboru je sice mnohem menší četnost konstruktů, ale i tak se klesající tendence projevila.

U Výběrového souboru 1 narušuje klesající tendenci pozorování $x_1 = 4$, tedy souvětí o délce 4 klauzí. Při bližším zkoumání jsme však neodhalili důvod tohoto odskoku. Hodnota koeficientu determinace je nízká $R^2 = 0{,}10$.

Data u Výběrového souboru 3 nejprve vykazují klesající tendenci (s odskokem v bodě $x_1 = 3$), avšak od souvětí složených ze sedmi a více klauzí $x_1 \geq 7$ se projevuje rostoucí tendence, a i když poslední tři body ($x_1 = 11$; $x_1 = 13$; $x_1 = 16$) mají opět klesající tendenci, jejich četnost je pouze $z_1 = 1$, proto je výsledná shoda nízká $R^2 = 0{,}19$.

Dále se podívejme, z jakého důvodu se u Výběrového souboru 2 projevila rostoucí tendence namísto klesající. Stejně jako na jazykové úrovni U2 i zde mohou výsledky ovlivňovat zejména přímé řeči s nedokončenou výpovědí, které patrně způsobují odchylku, což jsme pozorovali již na jazykové úrovni 2 (viz kapitola 4.2). Z celkového počtu 28 souvětí, jejichž délka $x_1 = 1$ (souvětí složené z jedné klauze), evidujeme 16 přímých řečí a 5 uvozovacích vět. Téměř ve všech případech, kdy je použita přímá řeč nebo uvozovací věta, se jedná o krátká souvětí, což je v rozporu s tvrzením MALu, které předpokládá, že nejkratší souvětí (tedy souvětí o délce 1 klauze), se budou skládat z nejdelších klauzí (měřených v průměrném počtu slov). Tendence je však zcela opačná, kupříkladu uvozovací věty typu *On/ona/... řekl(a)*. obsahují pouze 2–3 slova a vyskytly se dokonce čtyřikrát ze všech pěti případů uvozovacích vět (o délce jedné klauze):

Ona řekla. Tā shuō. 她说。

On řekl. Tā shuō. 他说。

Lin Hong řekla. Lín Hóng shuō. 林红说。

Li Hanlin řekl. Lǐ Hànlín shuō. 李汉林说。

Stejně tak u přímých řečí (souvětí o délce jedné klauze) jsou z celkového počtu 16 souvětí nejčetnější klauze složené z jednoho a třech slov (každé se vyskytlo celkem pětkrát), jedna klauze má délku 2 slova, delší klauze složené z 4, 5 a 6 slov se vyskytly každá pouze jednou a klauze o 8 slovech dvakrát.

Z dat získaných ze všech výběrových souborů můžeme prohlásit, že vztah mezi délkou konstruktů a jejich konstituentů, tedy čím více klauzí obsahují souvětí, tím méně slov obsahují v průměru dané klauze, se v rámci našeho výzkumu potvrdilo na nejvyšší jazykové úrovni U1 pouze u textů psaných latinkou. Je možné, že autoři, kteří píší text přímo v latince, při psaní podvědomě více sledují tendence, které definuje MAL. Oproti výběrovým souborům, které jsou psané ve znacích, jsou totiž v textech psaných přímo latinkou použity mnohem kratší souvětí (v počtu klauzí) a autoři pravděpodobně nevědomky zohledňují i stránku grafickou, protože pinyin určitým způsobem viditelně odráží zvukovou stránku jazyka a délku vyslovených slov, což čínské znakové písmo na první pohled neumí.

3.3.1. Experiment 3

Abychom postupovali systematicky, také na této jazykové úrovni prověříme, zda námi navržený způsob segmentace začleňující funkční slova jako součást jednoho celku (viz 3.1.1 Experiment 1) bude lépe odpovídá předpokladům MALu oproti původnímu způsobu segmentace. Slovo zde vystupuje jako jednotka, ve které jsou měřeny průměrné délky klauzí, proto délka slova stále může ovlivňovat celková data. Výsledky nám mohou napovědět, zda je alternativní způsob segmentace na slova vhodnější pro určení obecné definice čínského slova. Přepokládáme, že v případě Výběrového souboru 2 však nebude mít odlišný způsob segmentace žádný vliv, jelikož je struktura souboru natolik specifická, že se v tomto textu na nejvyšší jazykové úrovni U1 pravděpodobně neprojevují zákonitosti dané MALem.

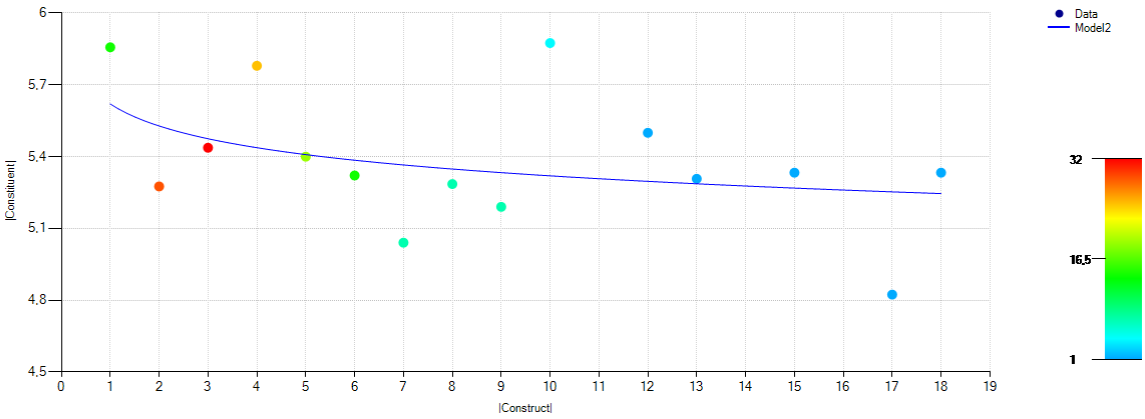
Následující Tabulka 31 zobrazuje data získaná po segmentaci a kvantifikaci textů po aplikovaní alternativního způsobu segmentace.

Tabulka 31 Experiment 3: Jazyková úroveň U1: souvětí (měřené v klauzích) – klauze (měřená v průměrném počtu slov)

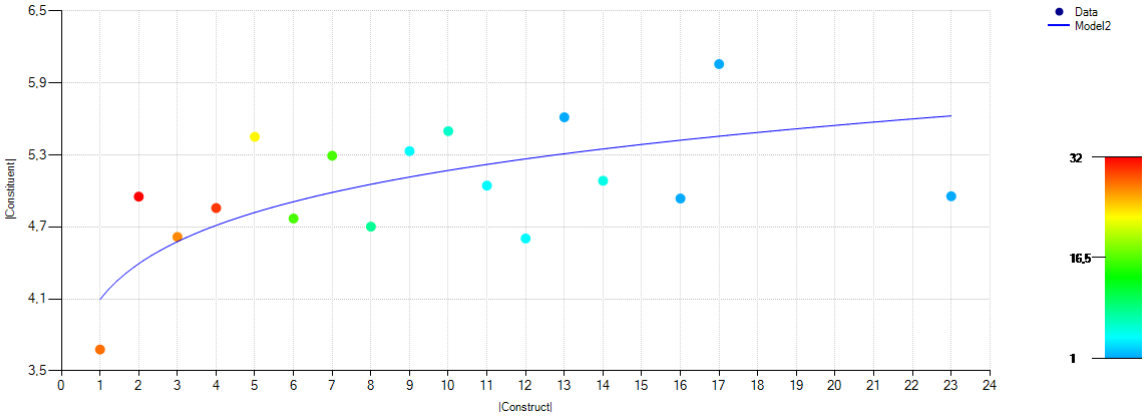
Výběrový soubor	**Délka souvětí v klauzích x_1**	**Frekvence z_1**	**Průměrná délka klauze ve slovech y_1**
Výběrový soubor 1	1	14	5,86
	2	28	5,28
	3	31	5,44
	4	25	5,78
	5	19	5,40
	6	14	5,32
	7	7	5,04
	8	7	5,29
	9	7	5,19
	10	4	5,88
	12	1	5,50
	13	1	5,31
	15	1	5,33
	17	1	4,82
	18	1	5,33
Výběrový soubor 2	1	28	3,68
	2	32	4,95
	3	27	4,62
	4	30	4,86
	5	23	5,45
	6	16	4,77
	7	16	5,29
	8	8	4,70
	9	4	5,33
	10	6	5,50
	11	4	5,05

	12	4	4,60
	13	1	5,62
	14	5	5,09
	16	1	4,94
	17	1	6,06
	23	1	4,96
Výběrový soubor 3	1	10	5,80
	2	8	4,75
	3	12	5,08
	4	9	4,67
	5	8	4,75
	6	7	4,52
	7	2	4,71
	8	2	5,06
	9	2	5,28
	10	2	6,40
	11	1	5,18
	13	1	5,69
	16	1	4,13

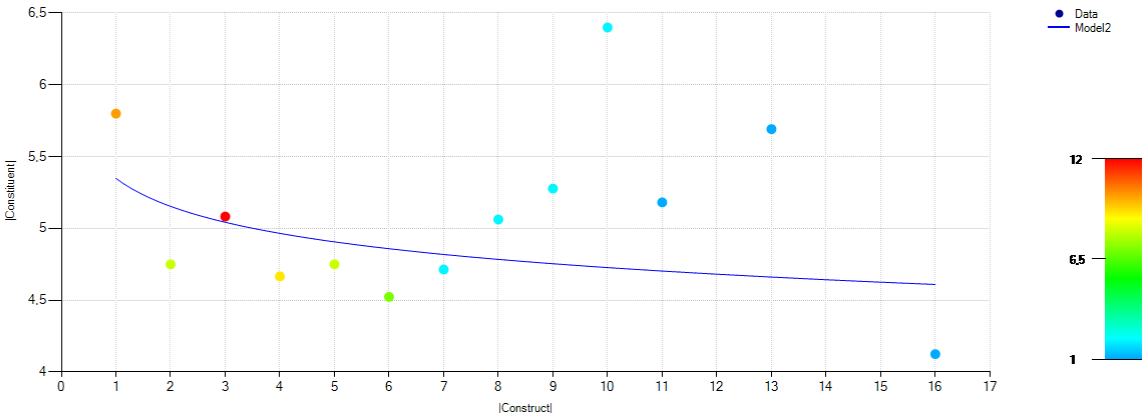
Alternativní způsob segmentace na této úrovni neovlivňuje délku souvětí v klauzích, jelikož jsme klauze segmentovali stále stejným způsobem, proto stavba textů zůstává stejná a délky souvětí jsou pro Výběrové soubory 1–3 opět od 1 do 23 klauzí. Avšak průměrná délka klauzí měřených ve slovech se nově pohybuje v intervalu $\langle 3{,}68; 6{,}70 \rangle$. Původně byla velikost intervalu $\langle 4{,}18; 7{,}40 \rangle$ a klauze tedy průměrně obsahovaly více slov. Z uvedených dat není na první pohled jasné, zda se bude projevovat klesající tendence, proto se podívejme na grafické vizualizace, viz Obrázek 38 – Obrázek 40.



Obrázek 38 Experiment 3: Grafická vizualizace pozorování uvedených v Tabulce 31 s proloženou regresní křivkou, Výběrový soubor 1: Yu Hua – *Kamarádi* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřená v průměrném počtu slov) pro námi navržený způsob segmentace



Obrázek 39 Experiment 3: Grafická vizualizace pozorování uvedených v Tabulce 31 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřená v průměrném počtu slov) pro námi navržený způsob segmentace



Obrázek 40 Experiment 3: Grafická vizualizace pozorování uvedených v Tabulce 31 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřené v průměrném počtu slov) pro námi navržený způsob segmentace

Klesající tendence jsou naznačeny u Výběrového souboru 1 a 3. Dle předpokladů se klesající tendence neprojevila u Výběrového souboru 2. I když grafické vizualizace víceméně zobrazují podobné tendence jako u původního způsobu segmentace, hodnoty koeficientu determinace se u Výběrového souboru 1 výrazně navýšily, z původní hodnoty $R^2 = 0{,}10$ jsme získali hodnotu $R^2 = 0{,}68$. Tento výsledek naznačuje, že alternativní způsob segmentace je vhodnější než původní způsob členění textu dle normy. U Výběrového souboru 2 má parametr *b* stále zápornou hodnotu a křivka má tedy rostoucí tendenci, z toho důvodu hodnota parametru R^2 není relevantní. Dle předpokladu se ani po aplikování alternativního způsobu segmentace tendence MALu nepotvrdila. U výběrového souboru 3 se hodnota koeficientu determinace příliš nezměnila, z původní hodnoty $R^2 = 0{,}19$ se snížila na $R^2 = 0{,}17$. Alternativní způsob segmentace tedy nemá na výsledky u tohoto výběrového souboru zásadní vliv.

Všechny hodnoty parametrů A a b i koeficientů determinace jsou přehledně uvedeny v následující Tabulce 32.

Tabulka 32 Experiment 3: Parametry *A* a *b* a koeficient determinace R^2 pro matematický model vztahující se k empiricky získaným datům na úrovni U1

Výběrový soubor	**Parametr *A***	**Parametr *b***	***Koeficient determinace R^2***
Výběrový soubor 1	5,62	0,02	0,68
Výběrový soubor 2	4,10	-0,10	0,54
Výběrový soubor 3	5,35	0,05	0,17

Shrnutí

Na nejvyšší jazykové hladině U1 souvětí – klauze se tendence definované MALem po segmentaci textů dle normy GB/T 16159–2012 projevily pouze u výběrových souborů psaných originálně v pinyinu. U Výběrového souboru 1 a 3 jsme sice pozorovali klesající tendenci předpokládanou MALem, ale výsledná shoda byla v obou případech nízká, a proto tvrzení, že slovo určené dle normy GB/T 16159–2012 odpovídá ekonomizujícím zákonům jazyka a odráží tak obecnou definici čínského slova, nelze potvrdit. U Výběrového souboru 2 se tendence definovaná MALem vůbec neprojevila, což je dle našeho názoru způsobeno specifickou strukturou tohoto výběrového souboru. Povídka obsahuje velké množství krátkých souvětí složených z jedné klauze, zejména přímé řeči a uvozovací věty. Předpokládáme, že právě struktura tohoto textu může mít vliv na celkové výsledky a způsobovat opačnou tendenci.

Stejně jako u předchozích jazykových úrovní jsme Výběrové soubory 1–3 segmentovali dle alternativního způsobu segmentace, který začleňuje vybraná funkční slova k plnovýznamovým slovům. Výrazně lepší výsledky jsme získali u Výběrového souboru 1, u kterého se po aplikování alternativního způsobu segmentace výrazně navýšila shoda s matematickým modelem MALu. U Výběrového souboru 2 dle předpokladů nedošlo ke zlepšení výsledků, kvůli výše zmíněné struktuře textu. Překvapivě u Výběrového souboru 3 neměl způsob segmentace na celkové výsledky téměř žádný vliv a shoda zůstává stále nízká, proto nelze říct, jestli slovo definované jedním či druhým způsobem více odpovídá ekonomizujícím pravidlům jazyka.

Po aplikování alternativního způsobu segmentace jsme získali lepší výsledky pouze u jednoho výběrového souboru, a proto se v dalším výzkumu zaměříme na zkoumání více různých typů výběrových souborů různého žánrového zaměření.

Dalším důvodem, proč se na této jazykové hladině neprojevil MAL, může být krátká délka zvolených výběrových souborů a z toho plynoucí nízká četnost jednotlivých konstruktů.

Jak jsme již zmiňovali na předchozí jazykové úrovni U2, jako další výzkum se nabízí segmentovat konstituent této jazykové úrovně, tj. klauzi, výhradně dle grafického principu, tzn. zohlednit pouze interpunkční znaménka čárku a středník jako hranici klauze v souvětí, což by ovlivnilo výsledky i na jazykové úrovni U1.

Závěr

Náš výzkum spadá do oblasti kvantitativní lingvistiky a jeho cílem bylo ověřit hypotézy, které jsme formulovali na základě výsledků získaných z předchozích experimentů. Z těchto výzkumů vyplynulo že, i když v textech psaných čínskými znaky chybí zřetelně vyznačené slovo, se slovem je při kvantitativních analýzách třeba počítat.

Předkládaná práce se zaměřila na ortografické slovo, které vznikne po převedení čínského textu dle normy GB/T 16159–2012 do abecedy *Hanyu pinyin Fang'an* (dále jen pinyin) jako část textu vytyčena mezerami. Naše hypotéza stanovila, že čínské ortografické slovo, které vznikne mezi mezerami po převodu čínského textu psaného ve znacích do abecedy pinyin dle normy GB/T 16159-2012, je v souladu s ekonomizujícími pravidly jazyka a je možné ho považovat za obecně platnou definici čínského slova. Jako prostředek ověřování hypotézy jsme zvolili Menzerath-Altmannův zákon (MAL).

Ověřování hypotézy probíhalo na pěti výběrových souborech: tři povídky byly psané v čínských znacích, které jsme následně převedli dle normy GB/T 16159–2012 do pinyinu. Další dva texty (deníkový záznam a učebnicový text) byly originálně zaznamenané formou pinyinu a nadále je tedy nebylo třeba segmentovat. U těchto dvou výběrových souborů jsme slova považovali jako části textu oddělené mezerami (případně interpunkcí). Zjistili jsme, že způsob segmentace u textů, které byly zaznamenány v pinyinu, se ve všech aspektech neshoduje s pravidly normy GB/T 16159–2012. V případě deníkového záznamu se segmentování lišilo pouze v několika případech, kdy autorka začleňovala přípony, záložky a záporky k plnovýznamovým slovům. U učebnicových textů jsme sledovali opačný trend a autoři naopak používali členitější slova.

Aby bylo možné ověřit ortografické slovo prostřednictvím MALu, který zkoumá vzájemný vztah dvou entit, bylo nutné definovat další sousední jazykové jednotky na nižších i vyšších jazykových hladinách. Vymezení jazykových jednotek *grafém* < *slabika* < ***slovo*** < *klauze* < *souvětí* vycházelo ze struktury psaných čínských textů a kombinovalo psanou podobu čínského znakového písma a převod znakového písma do latinky. Zmíněné jazykové jednotky byly spojeny do vzájemných vtahů a vznikly tak tři jazykové úrovně.

Největší pozornost byla věnována nejnižší jazykové úrovni U3, na které jsme ověřovali slovo na pozici konstruktu měřeného v průměrném počtu slabik. Po segmentaci textu na slova dle normy GB/T 16159–2012 jsme pomocí MALu zjistili, že ortografické

slovo získané tímto způsobem neodpovídá předpokladům MALu, a to ani u jednoho z výběrových souborů. Pravděpodobným důvodem, proč se neprokázaly ekonomizující zákony, byla vysoká frekvence jednoslabičných krátkých slov. Z toho důvodu jsme se rozhodli pro alternativní způsob segmentace, který vycházel jednak ze způsobu segmentace textu psaného latinkou (deníkový záznam), jednak z variantních zápisů, které byly obsaženy ve zmíněné normě. Námi navržený způsob segmentace začleňoval vybraná jednoslabičná funkční slova k plnovýznamovým slovům. Tímto způsobem jsme segmentovali povídky, které byly původně psané ve znacích. Výsledky získané z těchto tří výběrových souborů naznačují, že alternativní způsob segmentace více vyhovuje ekonomizujícím pravidlům v jazyce, což jsme porovnali ve dvou krocích. Nejprve jsme povídky segmentovali alternativním způsobem a u jedné z povídek jsme získali výsledky, které potvrdily, že slovo definované tímto způsobem je ve shodě s ekonomizujícími zákony jazyka. U dalších dvou povídek se shoda sice neprokázala, ale tendence se z jasně rostoucí začala pomalu převracet. Dále jsme u těchto dvou povídek odebrali nejčetnější jednoslabičná osobní zájmena jak u alternativního způsobu segmentace, tak i u původního způsobu segmentace, a zjistili jsme, že pouze u alternativního způsobu segmentace texty sledují tendenci MALu.

Jelikož MAL umožňuje jazykové jednotky ověřovat na různých hladinách, dále jsme se zabývali jazykovou hladinou U2, v rámci které stálo slovo na pozici konstituentu a konstruktem byla klauze měřená v průměrném počtu slov. Výsledky získané na této jazykové úrovni nebyly jednoznačné. Pouze výsledky ze dvou výběrových souborů (povídka *Kamarádi* a deníkový záznam psaný originálně v pinyinu) potvrdily hypotézu, že ortografické slovo vyhovuje ekonomickým pravidlům jazyka. Přičemž nejlepší data jsme získali u textu originálně psaného v pinyinu. U ostatních výběrových souborů se buď shoda neprokázala v dostatečné výši (v případě povídek), nebo se neprojevila vůbec (učebnicový text). Z důvodu získání rozporuplných výsledků nelze na této jazykové úrovni jednoznačně potvrdit hypotézu, že ortografické slovo určené dle normy se řídí ekonomizujícími zákony. Aby tato hypotéza mohla být potvrzena, je nutné ji ověřit na větším vzorku výběrových souborů.

Jelikož slovo na této jazykové úrovni určuje délku klauzí, přistoupili jsme i na této jazykové úrovni k alternativnímu způsobu segmentace, který začleňuje vybraná funkční slova k slovům plnovýznamovým a testovali jsme, zda alternativní způsob segmentace přinese lepší výsledky. Tento předpoklad se však na této jazykové úrovni nepotvrdil a ve všech případech výsledné procentuální shody dosáhly minimálních hodnot. Možným

důvodem nesouladu s ekonomizující pravidly je, že alternativní způsob segmentace není pro členění textu vhodnější, což je však třeba nadále ověřovat na dalších výběrových souborech. Mimoto jsme však zjistili, že platnost MALu výrazným způsobem narušuje stavba zvolených výběrových souborů. Některé soubory obsahovaly přímé řeči s nedokončenými výpověďmi, a právě ony by mohly narušovat ekonomičnost jazyka. Důvodem nesouladu s ekonomizující pravidly jazyka může být také nevyhovující členění textu na klauze. V jiných lingvistických pracích o čínštině se můžeme setkat s definováním klauze jako jazykové jednotky, která je v souvětí vymezena výhradně dle použité interpunkce a její hranici tvoří čárky a středníky.

Nejvyšší jazykové hladině U1 souvětí – klauze jsme nevěnovali tolik pozornosti, jako předchozím dvěma nižším jazykovým hladinám, protože primárně nebyla předmětem našeho výzkumu. A to jednat z důvodu, že slovo na této hladině sloužilo pouze k výpočtům průměrných délek klauzí, a také proto, že na této jazykové úrovni byl získán nižší počet empirických dat v porovnání s ostatními jazykovými úrovněmi. Nejlepší výsledky jsme na této jazykové hladině získali u výběrových souborů psaných originálně v pinyinu, u dalších dvou povídek byly výsledné shody minimální a u jedné povídky se projevila dokonce opačná tendence, což je pravděpodobně způsobeno specifickou strukturou tohoto výběrového souboru, který obsahuje velké množství krátkých souvětí složených z jedné klauze (zejména přímé řeči a uvozovací věty).

Z důvodu koherence jsme i na této jazykové úrovni aplikovali na výběrové soubory psané ve znacích alternativní způsob segmentace, který začleňuje vybraná jednoslabičná funkční slova k plnovýznamovým. Výrazně lepší výsledky jsme získali pouze u jedné povídky, u další povídky zůstala shoda stále minimální a u poslední povídky, která vykazovala opačnou tendenci již u původního způsobu segmentování, se tendence nezměnila. To má na svědomí pravděpodobně výše zmíněná specifická struktura textu. Členění textu na slova dle alternativního způsobu segmentace přineslo lepší výsledky pouze v jednom případě, proto opět nelze potvrdit, že ortografická slova, která v textu vzniknou po aplikování námi navrženého způsobu segmentace, odpovídají ekonomizujícím pravidlům v jazyce a potažmo obecné definici čínského slova, i když se tento způsob segmentace zdá být vhodnější. Na závěr neopomeňme zmínit, že výsledky získané na této jazykové úrovni mohou být zkreslené kvůli nedostatečné četnosti jednotlivých konstruktů.

Pokud porovnáme data získaná po kvantifikaci všech výběrových souborů, nejlepší výsledky (nejvyšší hodnoty koeficientů determinace) jsme na jazykových

úrovních U2 a U1 získali u deníkového záznamu psaného přímo v latince. Bez jakýchkoli zásahů do stavby textu dosahovaly hodnoty koeficientů determinace velmi vysokých hodnot. Důvodem pro získání lepších dat může být způsob členění textu na slova, který v některých případech začleňoval přípony, záložky a záporky k plnovýznamovým slovům. Dalším důvodem může být, že autorka při psaní zohledňuje estetiku psaného textu, kdy na první pohled vidí délku čínských slov a přizpůsobuje tomu i délky klauzí, potažmo souvětí (naopak při psaní vět v čínských znacích není délka vyslovených slov na první pohled patrná a autoři tolik nezohledňují zvukovou stránku, kterou určitým způsobem odráží pinyin). Je samozřejmě také možné, že konkrétně tato autorka při psaní sleduje tendence definované MALem.

Hypotéza, že ortografické slovo definované dle normy GB/T 16159–2012 odpovídá ekonomizujícím zákonům formulovaných MALem a odráží tak obecně platnou definici čínského slova nebyla na základě zkoumání zvolených výběrových souborů potvrzena, avšak nelze ji ani zamítnout, protože v průběhu výzkumu jsme zjistili, že platnost MALu může být ovlivněna různými faktory vycházející ze stavby zkoumaných textů, které mohou narušovat ekonomičnost. Prvním významným faktorem ovlivňující platnost MALu se zdá být vysoká frekvence pozorování, dále MAL reaguje velmi dynamicky na různé alternace textových produkcí, například když vnášíme přímou řeč, uvozovací věty, nahrazujeme osobní jména zájmeny, či sledujeme stylistické, pragmatické nebo estetické požadavky. Některé aspekty mohou jít ruku v ruce s ekonomičností jazyka, ale dle získaných výsledků pozorujeme, že některé jdou proti ní. Proto je nutné klást si otázku, je MAL opravdu tím nejideálnějším nástrojem pro ověření jazykové jednotky? Dle našeho názoru je možné ho pro ověření jazykové jednotky použít, avšak pro získání prokazatelnějších výsledků by se do výzkumu mělo kromě MALu zapojit více kvantitativně-lingvistických metod, které by nebyly tolik ovlivněné zmíněnými specifiky výběrových souborů. Z toho důvodu je třeba ve výzkumu nadále pokračovat.

Přestože získané výsledky nemůžeme generalizovat, práce poskytuje některé významné implikace. Na základě provedených experimentů jsme zjistili, že definice slova, která by více odpovídala předpokladům MALu, by měla zahrnovat spojování funkčních slov se slovy plnovýznamovými, což vychází jednak z provedených experimentů v rámci jazykové úrovně U3 a U1, ale také ze způsobu segmentace deníkového záznamu psaného originálně v pinyinu, který do určité míry funkční slova připojoval a u kterého jsme získali

nejlepší výsledky na jazykových úrovních U2 a U1. Tento způsob segmentace na slova je také nadále třeba ověřovat.

Návrhy na další výzkum

Závěrem uvádíme souhrn experimentů, kterými se budeme zabývat v navazujícím výzkumu:

1) Analýza textů psaných v pinyinu: v navazujícím výzkumu se zaměříme na analýzu více textů psaných originálně v pinyinu, protože právě u tohoto typu výběrového souboru jsme získali nejlepší výsledky. Budeme zjišťovat, jestli členění textů na slova odpovídá i v ostatních výběrových souborech od různých autorů psaných v pinyinu ekonomickým pravidlům jazyka.
2) Analýza výběrových souborů jiného žánrového zaměření: další výzkum bude zkoumat větší vzorek výběrových souborů jiného žánrového zaměření, na který budou aplikovány oba způsoby segmentace (vycházející z normy GB/T 16159–2012, a také z námi navrženého způsobu segmentace začleňující vybraná funkční slova k slovům plnovýznamovým)
3) Segmentace klauzí dle interpunkce: v jiných lingvistických pracích o čínštině se můžeme setkat s vymezením klauze jako jazykové jednotky, která je určena čistě graficky podle čárek a středníků v souvětí. Proto v dalším výzkumu budeme zjišťovat, jaký vliv na slovo má odlišné segmentování jazykové jednotky klauze (na jazykových úrovních U2 klauze – slovo i U1 souvětí – klauze).
4) Použití dalších kvantitativně-lingvistických zákonů: stanovené hypotézy budeme nadále ověřovat nejenom prostřednictvím MALu, ale také pomocí dalších kvantitativně-lingvistických metod (např. Brevity law apod.)
5) Přístupy k čínskému slovu: výzkum je možné rozšířit na ověřování ekonomických pravidel u slov definovaných odlišným způsobem (např. syntaktické, lexikální, fonetické slovo ad.)

Resumé

In the Chinese written texts, characters are strung together one after another without spaces. Thus, word boundaries are not indicated in any way in Chinese texts. The present paper focuses on Chinese orthographic word which is delimited by spaces after transcribing texts written in Chinese characters into Pinyin alphabet (*Hanyu Pinyin Fang'an*) according to the standard GB/T 16159–2012. The research tries to verify whether the word conforms to the economizing rules of the language. The hypothesis states that a Chinese orthographic word is in accordance with language economy, therefore the orthographic word delimited by this way can be considered as a general definition of the Chinese word. The orthographic word is verified by means of the Menzerath-Altmann Law which is quantitative linguistics' statistical method. Three language levels were acquired U3 *word* > *syllable* > *grapheme*, U2 *clause* > *word* > *syllable* and U1 *sentence* > *clause* > *word*. The hypothesis was tested on five different samples: three short stories written in Chinese characters and two samples originally written in pinyin. Results show that the orthographic word defined by the standard GB/T 16159–2012 does not conform to the economizing rules, thus the hypothesis was not confirmed. Significantly more appropriate results were obtained from one of the sample texts written in pinyin, which implies that the segmentation of the Chinese word should also combine selected function words (grammatical words) with full-meaning words so that the economizing rules are followed. Furthermore, the results indicate that written text are not always in compliance with the language economy and the validity of Menzerath-Altmann law can be influenced by factors such as usage of reporting and reported clauses, when personal names are replaced by pronouns and finally, it seems that the frequency also has a great impact on the validity of the Menzerath-Altmann law. Suggestions for further research are presented at the end of this thesis.

Key words

Quantitative linguistics, Menzerath-Altmann Law, written Chinese, Hanyu pinyin fang'an, orthographic word, language units

Literatura

Monografie a články

ALTMANN, Gabriel (1980). Prolegomena to Menzerath's Law. In: *Glottometrika*, 2, s. 124–129.

ALTMANN, Gabriel a Peter MEYER (2005). Physicist's look at language. In: Gabriel Altmann, Viktor Levickij a Valentina Perebyinis (ed.): *Problems of Quantitative* Linguistics. Černivci, s. 42–59.

ANDRES, Jan (2010). On a Conjecture about the Fractal Structure of Language. In: *Journal of Quantitative Linguistics*, 17 (2), s. 101–122.

ANDRES, Jan, Lubomír KUBÁČEK, Jitka MACHALOVÁ a Michaela TUČKOVÁ (2012). Optimization of parameters in the Menzerath–Altmann law. In: *Acta Univ. Palacki. Olomuc.*, Fac. Rer. Nat., Mathematica 51, 1, s. 5–27.

ANDRES, Jan, Martina BENEŠOVÁ, Martina CHVOSTEKOVÁ a Eva FIŠEROVÁ (2014). Optimization of parameters in the Menzerath–Altmann law, II. In: *Acta Univ. Palacki. Olomuc.*, Fac. Rer. Nat., Mathematica 53, 1, s. 3–25.

ANDRES, Jan, BENEŠOVÁ, Martina a LANGER, Jiří (2019). Towards a Fractal Analysis of the Sign Language. In: *Journal of Quantitative Linguistics*, 09.

BAXTER, William H. (1992). *A Handbook of Old Chinese Phonology*. Berlin, New York: Mouton De Gruyter. (Trends in Linguistics. Studies and Monographs 64), 935 s. ISBN 9783110123241

BENEŠOVÁ, Martina (2011). *Kvantitativní analýza textu se zvláštním zřetelem k analýze fraktální*. Disertační práce. Olomouc: Univerzita Palackého v Olomouci.

BENTZ, Christian a Ramon FERRER-I-CANCHO (2016). Zipf's law of abbreviation as a language universal. In: Bentz, Christian, Gerhard Jäger a Igor Yanovich (eds.) *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University of Tübingen, online publication system, https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558.

BOHN, Hartmut. *Quantitative Untersuchungen der modernen chinesischen Sprache und Schrift*. HAMBURG: DR. KOVAČ, 1998.

BORODA, Moisei G. A Gabriel ALTMANN (1991). Menzerath's Law In Musical Texts. In: *Musikometrica*, 3, s. 1–13.

BREITER, Maria A. (1994). Length of Chinese Words in Relation to Their Other Systemic Features. In: J. *Quant. Linguistics*, 1 (3), s. 224–231.

BUDÍKOVÁ, Marie, Maria KRÁLOVÁ a Bohumil MAROŠ (2010). *Průvodce základními statistickými metodami*. Praha: Grada. Expert. ISBN 978-80-247-3243-5.

CLINK DJ, Lau AR. Adherence to Menzerath's Law is the exception (not the rule) in three duetting primate species (2020). In: *Royal Society open science*. 7 (11):201557. https://doi.org/10.1098/rsos.201557 PMID:33391812

DeFRANCIS, John (1984). *The Chinese Language: Fact and Fantasy*. Honolulu: University of Hawaii Press. 331 s. ISBN 0-8284-0866-5

FENG, Zhiwei (1995). Language Technology and Language Resources in China. In: *Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure)* European Seminar (1st, Tihany, Hungary, September 15–16, s. 21–35.

FENG, Zhiwei 冯志伟 (2012): Yòng jìliàng fāngfǎ yánjiū yǔyán《用计量方法研究语言》 [Studium jazyka kvantitativními metodami]. In: *Foreign Language Teaching and Research (bimonthly) 《外语教学与研究（外国语文双月刊）》*). Vol. 44, No. 2.

FENG, Zhiwei a YIN Binyong (2000). The Chinese Digraphia Problem in the Information age. In: *Studies in the Linguistic Sciences*, Vol. 30, No. 1.

FERRER-I-CANCHO, Ramon a FORNS, Núria (2009). The self-organization of genomes. In: *Complexity*. 15 (5), s. 34–36.

FERRER-I-CANCHO, Ramon, Antoni HERNÁNDEZ-FERNÁNDEZ, Jaume BAIXERIES, Łukasz DĘBOWSKI a Ján MAČUTEK (2014). When is Menzerath-Altmann law mathematically trivial? A new approach. In: *Statistical Applications in Genetics and Molecular Biologyi*, 13 (6): s. 633–644.

FU, Huaiqing 符淮青 (2020). Xiàndài hànyǔ cíhuì (Zhòng pái běn) *《现代汉语词汇 （重排本）》* [*Slovní zásoba moderní čínštiny (přepracované vydání)*]. Beijing: Beijing daxue chubanshe. ISBN 978-7-301-30729-8.

GAN, Kok-Wee, Martha PALMER a Kim-Teng LUA (1996). A Statistically Emergent Approach for Language Processing: Application to Modeling Context Effects in Ambiguous Chinese Word Boundary Perception. In: *Computational Linguistics*, Vol. 22, No. 4, December. Dostupné z: https://aclanthology.org/J96-4004/

HANNAS, William C. (1997). *Asia's Orthographic Dilemma*. University of Hawai'i Press. ISBN 978-0-8248-1892-0.

HOU, Renkui, HUANG Chu-Ren, Hue San DO a Hongchao LIU (2017). A Study on Correlation between Chinese Sentence and Constituting Clauses Based on the Menzerath-Altmann Law. In: *Journal of Quantitative Linguistics*, ISSN: 1744-5035, DOI: 10.1080/09296174.2017.1314411

HOU, Renkui, HUANG Chu-Ren, AHRENS Kathleen a LEE Yat-Mei Sophia (2019). Linguistic characteristics of Chinese register based on the Menzerath—Altmann law and text clustering. In: *Digital Scholarship in the Humanities*, 02.

HŘEBÍČEK, Luděk (2002). *Vyprávění o lingvistických experimentech s textem*. Praha: Academia. 195 s. ISBN 80-200-0973-6.

HŘEBÍČEK, Luděk (2007). *Text in semantics: the principle of compositeness*. Praque: Academy of Sciences of the Czech Republic, Oriental Institute. Ex Oriente. ISBN 978-80-85425-59-8.

HU, Qianli (2005). On Chinese Romanization and Syllable Aggregation. In: *Cataloging & Classification Quarterly*, 40 (2), s. 19-32, DOI: 10.1300/J104v40n02_04

HUANG, Chu-Ren a XUE Nianwen (2012). Words without Boundaries: Computational Approaches to Chinese Word Segmentation. In: *Language and Linguistics Compass*, Vol. 6, Iss. 8, s. 494–505.

HUANG, Chu-Ren, Petr ŠIMON, Shu-Kai HSIEH, Laurent PRÉVOT (2007). Rethinking chinese word segmentation: tokenization, character classification, or wordbreak identification. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, s. 69–72.

HUANG, Chu-Ren, Ting-Shuo YO, Petr SIMON, Shu-Kai HSIEH (2008). A Realistic and Robust Model for Chinese Word Segmentation. Konference: *Proceedings of the*

20th Conference on Computational Linguistics and Speech Processing. Taipei, Taiwan, September 4–5. Association for Computational Linguistics.

CHAPPELL, Hilary (1980). The Romanization Debate. In. *The Australian Journal of Chinese Affairs*, No. 4, s. 105–118.

CHEN, Heqin 陈鹤琴 (1928). *Yutiwen yingyong zihui 《语体文应用字汇》 [Aplikovaná slovní zásoba hovorových textů]*. Shanghai: Shangwu yinshuguan 《商务印书馆》, 117 s. Dostupné z: https://taiwanebook.ncl.edu.tw/zh-tw/book/NCL-000498284/reader

CHEN, Heng a LIU Haitao (2014). A diachronic study of Chinese word length distribution. In: *Glottometrics*, 29, s. 81–94.

CHEN, Heng a LIU Haitao (2022). Approaching language levels and registers in written Chinese with the Menzerath–Altmann Law. In: *Digital Scholarship in the Humanities*. DOI: 10.1093/llc/fqab110

CHEN Heng, LIANG Junying a LIU Haitao (2015). How Does Word Length Evolve in Written Chinese? In: *PloS ONE* 10 (9), s. 1–12. Dostupné z: doi:10.1371/journal.pone.0138567

CHEN, Ping (1993). Modern Written Chinese in Development. In: *Language in Society*, Vol. 22, No. 4, s. 505–537. JSTOR, www.jstor.org/stable/4168472.

CHEN, Ping (1994). Four Projected Functions of New Writing Systems for Chinese. In: *Anthropological Linguistics*, Vol. 36, No. 3, s. 366–381. *JSTOR*, www.jstor.org/stable/30028029.

CHEN, Ping (1999). *Modern Chinese: history and sociolinguistics*. New York: Cambridge University Press. 229 s. ISBN 05-216-4572-7.

JOHNSON, Keith (2008). *Quantitative Methods in Linguistics*. Oxford: Blackwell. ISBN 978-1-4051-4424-7.

KANE, Daniel (2009). *Knížka o čínštině*. Mirošovice: Desert Rose. ISBN 978-80-903296-1-4.

KASKE, Elisabeth (2008). *The Politics of Language in Chinese Education: 1895 – 1919*. Leiden: BRILL, 537 s. ISBN 9004163670, 9789004163676

KÖHLER, Reinhard a Gabriel ALTMANN (2005). Aims and Methods of Quantitative Linguistics. In: *Problems of Quantitative Linguistics*, s. 12–41.

KUŁACKA, Agniezska (2010). The Coefficients in the Formula for the Menzerath-Altmann Law. In: *Journal of Quantitative Linguistics*. Vol. 17, No. 4, s. 257–268.

LANGER, Jiří et al. (2020). *Quantitative Linguistic Analysis of Czech Sign Language.* Odborné publikace 1. vydání, 2020, 216 s., ISBN 978-80-244-5728-4.

LE, Quan-Ha, Elvira I. SICILIA-GARCIA, JI Ming a F. Jack SMITH (2002). Extension of Zipf's law to words and phrases. In: *Proceedings of the 19th international 140em a140ence on computational linguistics* (COLING-2002), Taibei, o. S.

LE, Quan-Ha, Elvira I. SICILIA-GARCIA, JI Ming a F. Jack SMITH (2003), Extension of Zipf's law to word and character *n*-grams for English and Chinese. In: *Computational linguistics and Chinese language processing*, 8 (1), s. 77–102.

LE Quan-Ha, P. HANNA, D. W. STEWART a F Jack SMITH (2006). Reduced *n*-gram models for English and Chinese corpora. In: *Proceedings of the COLING/ACL on Main 140onference poster sessions*.

LI, Dejin, Meizhe CHENG a Dehou JIN (2008). *Waiguoren shiyong Hanyu yufa = A practical Chinese grammar for foreigners*. Xiu ding ben, di 1 ban. Beijing: Beijing yuyan daxue chubanshe. ISBN 978-7-5619-2163-0.

LI, Wentian (2012). Menzerath's law at the gene-exon level in the human genome. In: *Complexity*, 17 (4), s. 49–53.

LI, Shoushan a Chu-Ren HUANG (2009). Word Boundary Decision with CRF for Chinese Word Segmentation. In: *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, s. 726–732. Dostupné z: https://aclanthology.org/Y09-2034.pdf

LIU, Haitao 刘海涛 a FENG Zhiwei 冯志伟 (2007): Zìrán yǔyán chǔlǐ de gàilǜ pèijià móshì lǐlùn《自然语言处理的概率配价模式理论》[Teorie pravděpodobnostního valenčního modelu pro zpracování přirozeného jazyka]. In: *Linguistic Sciences 《语言科学》*, 6 (3), s. 32–41.

LIU, Haitao 刘海涛 a HUANG Wei 黄伟 (2012). Jìliàng yǔyánxué de xiànzhuàng, lǐlùn yǔ fāngfǎ《计量语言学的现状、理论与方法》

[Současná situace, teorie a metody kvantitativní lingvistiky]. In: *Journal of Zhejiang University (Humanities and Social Sciences)*. Vol. 42, No. 2.

LIU, Ping-Ping, LI, Wei-Jun, LIN, Nan, LI, Xing-Shan, PELLI, Denis G. (2013). Do Chinese Readers Follow the National Standard Rules for Word Segmentation during Reading? In: *PloS ONE*. 2 Vol. 8; Iss. 2.

LIU, Wei a Louise GUTHRIE (2009). Chinese Pinyin-Text Conversion on Segmented Text. Conference paper. In: *TSD 2009: Text, Speech and Dialogue*, s. 116–123.

LIU, Yongbing (2005). A Pedagogy for Digraphia: An Analysis of the Impact of Pinyin on Literacy Teaching in China and its Implications for Curricular and Pedagogical Innovations in a Wider Community. In: *Language and Education*, 19:5, s. 400–414, DOI: 10.1080/09500780508668693

LIU, Yuan 刘源 (1994). *Xìnxī chǔlǐ yòng xiàndài Hànyǔ fēncí guīfàn jí zìdòng fēncí fāngfǎ《信息处理用现代汉语分词规范及自动分词方法》[Norma pro segmentaci slov moderního čínského jazyka pro zpracování dat a způsob automatické segmentace slov]*. Qinghua daxue chubanshe. 474 s. ISBN 9787302014300.

LUA, Kim Teng (1995). Predication of Meaning of Bisyllabic Chinese Compound Words Using Back Propagation Neural Network. In: *Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation*, s. 49–56.

MAIR, Victor H. (2002). Sound and Meaning in the History of Characters: Views of China‘s Earliest Script Reformers. In: ERBAUGH, Mary S. ed. *Difficult Characters: Interdisciplinary Studies of Chinese and Japanese Writing*. Colombus, Ohio: Ohio State University National East Asian Language Resource Center. 285 s. ISBN-13: 978-0874153446. Dostupné z: http://pinyin.info/readings/difficult_characters.html

MATLACH, Vladimír, Daniel DOSTÁL a Marian NOVOTNÝ (2022). Secondary Structures of Proteins Follow Menzerath–Altmann Law. In: *Int. J. Mol. Sci*. 2022, 23, 1569. Dostupné z: https://doi.org/10.3390/ijms23031569

McCALLUM, Andrew a Fang-fang FENG (2003). *Chinese Word Segmentation with Conditional Random Fields and Integrated Domain Knowledge.* Dostupné z: https://people.cs.umass.edu/~mccallum/papers/chineseseg.pdf

MATOUŠKOVÁ, Lenka a Tereza MOTALOVÁ (2015). An Application of the Menzerath-Altmann law to Chinese translations of the poem The Raven. In: *Czech and Slovak Linguistic Review*. 2/2015. s. 44-61.

MATOUŠKOVÁ, Lenka (2016). Application of the Menzerath-Altmann Law to a Text Written in Traditional Chinese Characters. In: Benešová, Martina et al.: *Text Segmentation for Menzerath-Altmann Law Testing*. Olomouc: Univerzita Palackého v Olomouci, s. 44–71 ISBN 978-80-244-5112-1.

MOTALOVÁ, Tereza a Lenka SPÁČILOVÁ (MATOUŠKOVÁ) (2013). *Aplikace Menzerath-Altmannova zákona na současnou psanou čínštinu*. Diplomová práce. Olomouc: Univerzita Palackého v Olomouci, Filozofická fakulta. Vedoucí práce Mgr. Ondřej Kučera.

MOTALOVÁ, Tereza a Lenka SPÁČILOVÁ (MATOUŠKOVÁ) (2014). *An Application of the Menzerath-Altmann Law to Contemporary Written Chinese*. Olomouc: Nakladatelství UP.

MOTALOVÁ, Tereza a Lenka SPÁČILOVÁ (MATOUŠKOVÁ), Martina Benešová a Ondřej Kučera (2016). An Application of the Menzerath-Altmann Law to Contemporary Written Chinese. In: Martina Benešová (ed.): *Menzerath-Altmann Law Applied*. Olomouc: Nakladatelství UP, s. 87–120.

MOTALOVÁ, Tereza a Denisa SCHUSTEROVÁ (2016). Menzerath-Altmann Law – Analyses of Short Stories Written by Chinese Authors. In: Benešová, Martina et al.: *Text Segmentation for Menzerath-Altmann Law Testing*. Olomouc: Univerzita Palackého v Olomouci, s. 72–117 ISBN 978-80-244-5112-1.

MUSHANGWE, Herbert a Godfrey CHISONI (2015). A Critical Analysis of the Use of Pinyin as a Substitute of Chinese Characters. In: *Journal of Language Teaching and Research*, Vol. 6, No. 3, s. 685–694, May. ISSN 1798-4769. DOI: http://dx.doi.org/10.17507/jltr.0603.28

NISHIMOTO, Eiji (2003). Measuring and comparing the productivity of Mandarin Chinese suffixes. In: *Computational linguistics and Chinese language processing*, 8 (1), s. 49–76.

NORMAN, Jerrv (1988). *Chinese*. New York: Cambridge University Press. ISBN 978-0-521-29653-3.

PACKARD, Jerome (2000). *The Morphology of Chinese: A Linguistic and Cognitive Approach*. Cambridge University Press, ISBN 9781139431668.

PALÁT, Augustin (1999). Několik slov k počátkům transkripce čínštiny u nás. In: Třísková, Hana (ed.): *Transkripce čínštiny*. Praha: Česko-čínská společnost, s. 23–27.

PENG, Haiyun, Yukun MA, Soujanya PORIA, Yang LI, Erik CAMBRIA (2019). *Phonetic-enriched Text Representation for Chinese Sentiment Analysis with Reinforcement Learning*. arXiv:1901.07880 [cs.CL] https://doi.org/10.48550/arXiv.1901.07880

POPESCU, I-I, Sven NAUMANN, Emmerich KELIH, Andrij ROVENCHAK, Anja OVERBECK, Haruko SANADA, Reginald SMITH, Radek ČECH, Panchanan MOHANTY, Andrew WILSON a Gabriel ALTMANN (2013). Word length: aspects and languages. In: G Altmann & R Köhler (eds): *Issues in Quantitative Linguistics Vol. 3*. Studies in Quantitative Linguistics, Vol. 13, RAM-Verlag, Lüdenscheid, s. 224–281.

PETRŮ, Eduard (2000). *Úvod do studia literární vědy*. Olomouc: Rubico. ISBN 80-85839-44-X.

QIAO, Wei, SUN Maosong, Wolfgang MENZEL (2008). Statistical Properties of Overlapping Ambiguities in Chinese Word Segmentation and a Strategy for Their Disambiguation. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds) *Text, Speech and Dialogue*. TSD 2008. Lecture Notes in Computer Science, Vol. 5246. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-87391-4_24

QIAO Wei, SUN Maosong a Wolfgang MENZEL (2010). Chinese Word Frequency Approximation Based on Multitype Corpora. In: *Journal of Quantitative Linguistics*, 17:2, s. 142–166, DOI: 10.1080/09296171003643213

RASINGER, Sebastian M. (2013). *Quantitative Research in Linguistics*: *an introduction*. Second edition. London: Bloomsbury. eISBN: 978-1-4725-6697-3

ROHSENOW, John S. (2001). The present status of digraphia in China. In: *International Journal of the Sociology of Language*, Vol. 2001; Iss. 150

ROUSSEAU, Ronald a ZHANG Qiaoqiao (1992). Zipf's data on the frequency of Chinese words revisited. In: *Scientometrics*, 24 (2), s. 201–220.

SEHNAL, David (1999). Co je to vlastně pinyin? In: Třísková, Hana (ed.): *Transkripce čínštiny*. Praha: Česko-čínská společnost, s. 85–88.

SEHNAL, David (2006). Čínský jazyk nebo čínské jazyky?. In: *Britské listy* [online]. 8. 9. 2006 [cit. 2019-03-25]. ISSN 1213-1792. Dostupné z: https://legacy.blisty.cz/art/30211.html

SEHNAL, David (2002). Čínské znakové písmo, jeho povaha a vývoj. In: Černá, Z. a Lomová, O. (eds.): *Z myšlenek a představ Žluté země*. Brno: Moravské zemské muzeum.

SHIH, Chilin a Richard SPROAT (1996). Issues in Text-to-Speech Conversion for Mandarin. In: *Computational Linguistics and Chinese Language Processing*.

SHIH, Hsiu-Chuan, a Jenny W. HSU (2008). Hanyu Pinyin to be standard system in 2009. In: *Taipei Times* [online]. [cit. 2014-06-29]. Dostupné z: http://www.taipeitimes.com/News/taiwan/archives/2008/09/18/2003423528

SHTRIKMAN, Shmuel (1994). Some comments on Zipf's law for the Chinese language. In: *Journal of information science*, 20 (2), s. 142–143.

SLAVÍČKOVÁ, Jana (1988). O jednom výběru reklamních textů pro sémantický popis jazyka (na materiálu z deníku Morning Star). In: *Slovo a slovesnost*, Vol. 49, No. 3, s. 209–215.

SPROAT, Richard W., Chilin SHIH, William GALE a Nancy CHANG (1996). A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. In: *Computational Linguistics*, 22 (3), s. 377–404.

SU, Peicheng (2001). Digraphia: a strategy for Chinese characters for the twenty-first century. In: *International Journal of the Sociology of Language*, Vol. 2001 (01); Iss. 150.

SUN, Chaofen (2006). *Chinese: A Linguistic Introduction*. Cambridge University Press. ISBN 1139453645

SUN, Fengjie a Gustavo CAETANO-ANOLLÉS (2021). Menzerath-Altmann's Law of Syntax in RNA Accretion History. In: *Life (Basel)*. May 27; 11 (6): 489. doi: 10.3390/life11060489.

ŠVARNÝ, Oldřich a kol. (1993). *Gramatika hovorové čínštiny v příkladech 2a*. Filozofická fakulta Univerzity Komenského. ISBN 80-223-0534-0.

ŠVARNÝ, Oldřich (1999). Fonetické a fonologické srovnání *pchin-jinu* a české standardní transkripce. In: Třísková, Hana (ed.): *Transkripce čínštiny*. Praha: Česko-čínská společnost, s. 41–46.

ŠVARNÝ, Oldřich a David UHER (2001). *Hovorová čínština. Úvod do studia hovorové čínštiny*. Olomouc: Univerzita Palackého v Olomouci. ISBN 80-224-0298-X.

TĚŠITELOVÁ, Marie. *Kvantitativní lingvistika*. Praha: SPN, 1987.

TORRE, Iván G., Łukasz Dębowski a Antoni Hernández-Fernández (2021). Can Menzerath's law be a criterion of complexity in communication? In: *PLoS One*. Aug 20; 16 (8): e0256133. doi: 10.1371/journal.pone.0256133.

TŘÍSKOVÁ, Hana (1999a). O přepisu cizích slov. In: Třísková, Hana (ed.): *Transkripce čínštiny*. Praha: Česko-čínská společnost, s. 3–7.

TŘÍSKOVÁ, Hana (1999b). Fonetický zápis čínštiny. In: Třísková, Hana (ed.): *Transkripce čínštiny*. Praha: Česko-čínská společnost, s. 13–20.

TŘÍSKOVÁ, Hana (2010). *Segmentální struktura čínské slabiky*. Disertační práce. Praha: Univerzita Karlova v Praze, Filozofická fakulta. Vedoucí práce: PhDr. Zdenka Hermanová, CSc.

VOCHALA, Jaromír a Věna HRDLIČKOVÁ (1985). *Úvod do studia sinologie: část filologická*. Praha: SPN.

VOCHALA, Jaromír (1999). Čínský přepis *pinyin* a jeho užití v češtině. In: Třísková, Hana (ed.): *Transkripce čínštiny*. Praha: Česko-čínská společnost, s. 33–39.

WAN, Defu (2014). The History of Language Planning and Reform in China: A Critical Perspective. In: *Working Papers in Educational Linguistics*, Vol. 29, No. 2 Fall. Article 5, s. 65–79. Dostupné z: https://pdfs.semanticscholar.org/6682/b18cb1006473ffb7dedc7ff2fd322f12a0cc.pdf

WANG, Li 王力 (2004). *Hànyǔ shǐgǎo《汉语史稿》 [Historický nástin čínského jazyka]*. Beijing: Zhonghua shuju. 714 s. ISNB 7-101-04199-X

WANG, Lin a Radek Čech (2016). The impact of code-switching on the Menzerath-Altmann Law. In: *Glottometrics*, 35. RAM-Verlag, s. 22–27. ISSN 2625-8226

WANG, Lu (2013). Word length in Chinese. In: Köhler, R. a Altmann, G. (Eds): *Issues in Quantitative Linguistics 3*. Dedicated to Karl-Heinz Best on the occasion of his 70th birthday, s. 39–53. Lüdenscheid: RAM.

WIMMER, Gejza, Gabriel ALTMANN, Luděk HŘEBÍČEK, Slavomír ONDREJOVIČ a Soňa WIMMEROVÁ (2003). *Úvod do analýzy textov*. Bratislava: VEDA, vydavateľstvo Slovenskej akadémie vied . 344 s. ISBN 80–224–0756–9.

WRENN, James J. (1975). Popularization of Putonghua. In: *Journal of Chinese Linguistics*, Vol. 3, No. 2/3, s. 221–227. JSTOR, www.jstor.org/stable/23749877.

XU, Lirong a HE Lianzhen (2018). Is the Menzerath-Altmann Law Specific to Certain Languages in Certain Registers? In: *Journal of Quantitative Linguistics*. ISSN 1744-5035. DOI: 10.1080/09296174.2018.1532158

XUE, Nianwen (2003). Chinese word segmentation as character tagging. In: *International Journal of Computational Linguistics and Chinese Language Processing*, 8 (1), s. 29–48.

YIN, Binyong a Mary FELLEY (1990). *Chinese Romanization: Pronunciation and Orthography*. Beijing: Sinolingua. ISBN: 978-7800521485

Yuen Ren CHAO (赵元任) (2011). *A Grammar of Spoken Chinese*. Beijing: The Commercial Press. ISBN 978-7-100-08739-1

ZÁDRAPA, Lukáš a Michaela PEJČOCHOVÁ (2009). *Čínské písmo*. Praha: Academia. Orient. ISBN 978-80-200-1755-0.

ZHANG, Ziquan 张子泉 (2005). *Pǔtōnghuà jiàochéng 《普通话教程》[Příručka o současné čínštině]*. Beijing: Qinghua daxue chubanshe. 172 s. ISBN 7302107254, 9787302107255.

ZHAO, Shouhui a Richard B. BALDAUF Jr. (2008). *Planning Chinese Characters. Reaction, Evolution or Revolution?*. Springer. 419 s. ISBN 978-0-387-48576-8 (e-book).

ZHAO, Yiheng (1991). Yu Hua: Fiction as Subversion. In: *World Literature Today*, Vol. 65, No. 3, Contemporary Chinese Literature, s. 415–420.

ZHOU, Youguang 周有光 (1980). Xiàndài Hànzixué Fāfán《现代汉子学发凡》 [Úvod do studia čínského znakového písma]. In: *Yuwen xiandaihua*, (2), s. 94–103.

ZHOU, Youguang 周有光 (1992). *Zhōngguó yǔwén zònghéngtán《中国语文纵横谈》 [Volná diskuse o čínském mluveném a psaném jazyce]*. Beijing: Renmin jiaoyu chubanshe.

ZHU Jinyang a Karl-Heinz BEST (1992). Zum Wort im modernen Chinesisch. In: *Oriens Extremus,* Vol. 35, No. 1/2, s. 45–60.

ZHU Jinyang a Karl-Heinz BEST (1997), Zur Modellierung der Wortlängen im Chinesischen. In: *Glottometrika* 16 (Hrsg. Karl-Heinz Best). Trier: Wissenschaftl. Verlag Trier, s. 185–194.

ZHU Jinyang a Karl-Heinz BEST (1998), Wortlängenhäufigkeiten in chinesischen Kurzgeschichten. In: *Asian and African Studies* 7 (Bratislava), s. 45–51.

Výběrové soubory

HAN, Han 韩寒 (2012). Wǒ suǒ lǐjiě de shēnghuó《我所理解的生活》 [Život, jak mu rozumím]. In: *http://blog.sina.com.cn/twocold* [online]. [cit. 2013-04-16]. Dostupné z: http://blog.sina.com.cn/s/blog_4701280b0102e7er.html

YU, Hua 余华 (2005). Péngyou《朋友》[Kamarádi]. In: *Yu Hua Jing xuanji《余华精选集》 [Yu Hua, sborník povídek]*. Beijing: Beijing chubanshe 北京 出版社. s. 1–14. ISBN 978-7-5402-0304-7

YU, Hua 余华 (2005). Nǚrén de shènglì《女人的胜利》 [Vítězství ženy]. In: *Yu Hua Jing xuanji《余华精选集》[Yu Hua, sborník povídek]*. Beijing: Beijing chubanshe 北京 出版社. s. 312–330. ISBN 978-7-5402-0304-7

ZHANG, Liqing (2010). Mihuo. In: *Pīnyīn Rìjì Duǎnwén*. Banqiao Shi, Taiwan: Pinyin.Info. Dostupné z: http://www.pinyin.info/readings/pinyin_riji_duanwen/04_mihuo.html

LIU, Yuehua a Daozhong YAO (2009). *Integrated Chinese: [Zhongwen tingshuo du xie]*. 3rd ed. Boston: Cheng & Tsui. s. 267, 280, 298, a 308. ISBN 978-0-88727-638-5.

Normy

GB/T 13715–92. Contemporary Chinese language word segmentation specification for information processing 《信息处理用现代汉语分词规范》 (1992). Čínská lidová republika: Ministerstvo strojírenství a elektronického průmyslu Čínské lidové republiky (中华人民共和国机械电子工业部). Dostupné z: https://data.gcbz.org/data/content/GBT%2013715-1992%20%E4%BF%A1%E6%81%AF%E5%A4%84%E7%90%86%E7%94%A8%E7%8E%B0%E4%BB%A3%E6%B1%89%E8%AF%AD%E5%88%86%E8%AF%8D%E8%A7%84%E8%8C%83.pdf?st=yuZw7bjPYf6n-RnNKCOZ1w&e=1652922022

GB/T16159–1996. *Basic rules for Hanyu Pinyin orthography 《汉语拼音正词法基本规则》* (1996) Čínská lidová republika: State Education Commission of the PRC (国家教育委员会) a State Language and Literature Working Committee of the PRC (国家语言文字工作委员会).

GB/T16159–2012. *Basic rules of the Chinese phonetic alphabet orthography 《汉语拼音正词法基本规则》* (2012) Čínská lidová republika: Standardization Administration of China (中国国家标准化管理委员会) a General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China (中华人民共和国国家质量监督检验检疫总局).

Hanyu pinyin fang'an《汉语拼音方案》(1958). Beijing: Wenzi gaige chubanshe (文字改革出版社). Dostupné z: http://www.moe.gov.cn/ewebeditor/uploadfile/2015/03/02/20150302165814246.pdf

Webové stránky a příspěvky

Biaodian fuhao 《标点符号》 [Interpunkční znaménka] (© 2017). *Baidu baike 《百度百科》* [online]. [cit. 8.7.2017]. Dostupné z: http://baike.baidu.com/view/31516.htm

Cizu《词组》 [Slovní spojení] (© 2022). *Baidu baike 《百度百科》* [online]. [cit. 27.06.2022]. Dostupné z: https://baike.baidu.com/item/%E8%AF%8D%E7%BB%84/8481390

Corpus annotation (2022). *Lancaster.ac.uk* [online]. Copyright © 2022 Lancaster University. [cit. 30.06.2022]. Dostupné z: https://www.lancaster.ac.uk/fass/projects/corpus/ZCTC/annotation.htm

Editorial board (2022). *Tandfonline.com* [online]. London: Copyright © 2022 Informa UK Limited [cit. 2022-07-27]. Dostupné z: https://www.tandfonline.com/action/journalInformation?show=editorialBoard&journalCode=njql20

Glottometrics – About (2022). *Glottometrics.iqla.org* [online] Chicago: Copyright © 2022 Glottometrics • Chicago by Catch Themes [cit. 2022-07-27]. Dostupné z: https://glottometrics.iqla.org/about/

Li-ching Chang, 1936-2010. *Pinyin News: the blog of Pinyin.info* [online]. WordPress, 2010 [cit. 2017-10-27]. Dostupné z: http://pinyin.info/news/2010/li-ching-chang-1936-2010/

Feng Zhiwei 《*冯志伟*》 (©2021): (Jiàoyùbù yǔyán wénzì yìngyòng yánjiūsuǒ yánjiūyuán, 教育部语言文字应用研究所研究员) [online]. *Baidu baike* 《*百度百科*》, 2021 [cit. 2021-04-10]. Dostupné z: https://baike.baidu.com/item/%E5%86%AF%E5%BF%97%E4%BC%9F/52693

Xinying Chen 陈芯莹 (2022). *Yuyanxue.net* [online] © 2018 All rights reserved [cit. 2022-07-27]. Dostupné z: https://www.yuyanxue.net/

Yīmùliǎorán chūjiē de zhǔyào nèiróng, yīmùliǎorán chūjiē dǎo dú《一目了然初阶》的主要内容,《一目了然初阶》导读 [Hlavní obsah a průvodce Yimuliaoran chujie] (2019). In: *Pinshiwen.com* [online]. [cit. 2021-5-15]. Dostupné z: https://www.pinshiwen.com/cidian/zpjx/20190731166478.html

Software

Wenlin Institute, Inc. Wenlin Software for Learning Chinese [software]. Version 4.0.2. Wenlin Institute, Inc. Copyright © 1997–2011.

BENEŠOVÁ, Martina, CHVOSTEKOVÁ, Martina, MATLACH, Vladimír. *MA Studio* [software]. 2015 Univerzita Palackého v Olomouci.

Seznam tabulek

Seznam obrázků

Přílohy

1. Základní pravopisná pravidla pinyinu, norma GB/T 13715–92
2. Základní pravopisná pravidla čínské fonetické abecedy, norma GB/T 16159–2012
3. Grafické vizualizace, parametry A a b a koeficienty determinace R^2 pro všechny modely MALu

1. Základní pravopisná pravidla pinyinu, norma GB/T 13715–92

中华人民共和国国家标准

汉语拼音正词法基本规则

GB/T 16159—1996

Basic rules for Hanyu Pinyin Orthography

1 主题内容与适用范围

本标准规定了用《汉语拼音方案》拼写现代汉语的规则。内容包括分词连写法、成语拼写法、外来词拼写法、人名地名拼写法、标调法、移行规则等。为了适应特殊的需要，同时提出一些可供技术处理的变通方式。

本标准适用于文教、出版、信息处理及其他部门，作为用《汉语拼音方案》拼写现代汉语的统一规范。

2 术语

汉语拼音正词法

汉语拼音的拼写规范及其书写格式的准则。《汉语拼音方案》确定了音节的拼写规则。《汉语拼音正词法基本规则》是在《汉语拼音方案》的基础上进一步规定词的拼写规范的基本要点。

3 制定原则

3.1 以词为拼写单位，并适当考虑语音、语义等因素，同时考虑词形长短适度。

3.2 基本采取按语法词类分节叙述。

3.3 规则条目尽可能详简适中，便于掌握应用。

4 汉语拼音正词法基本规则

4.1 总原则

4.1.1 拼写普通话基本上以词为书写单位。

rén(人)　pǎo(跑)　hǎo(好)　hé(和)　hěn(很)

fúróng(芙蓉)	qiǎokèlì(巧克力)
péngyou(朋友)	yuèdú(阅读)
dìzhèn(地震)	niánqīng(年轻)
zhòngshì(重视)	wǎnhuì(晚会)
qiānmíng(签名)	shìwēi(示威)
niǔzhuǎn(扭转)	chuánzhī(船只)
dànshì(但是)	fēicháng(非常)
diànshìjī(电视机)	túshūguǎn(图书馆)

4.1.2 表示一个整体概念的双音节和三音节结构，连写。

gāngtiě(钢铁)	wèndá(问答)
hǎifēng(海风)	hóngqí(红旗)
dàhuì(大会)	quánguó(全国)

国家技术监督局1996-01-22批准　　1996-07-01实施

zhòngtián(种田)　　kāihuì(开会)
dǎpò(打破)　　zǒulái(走来)
húshuō(胡说)　　dǎnxiǎo(胆小)
qiūhǎitáng(秋海棠)　　àiniǎozhōu(爱鸟周)
duìbuqǐ(对不起)　　chīdexiāo(吃得消)

4.1.3 四音节以上表示一个整体概念的名称，按词(或语节)分开写，不能按词(或语节)划分的，全都连写。

wúfèng gāngguǎn(无缝钢管)
huánjìng bǎohù guīhuà(环境保护规划)
jīngtǐguǎn gōnglǜ fàngdàqì(晶体管功率放大器)
Zhōnghuá Rénmín Gònghéguó(中华人民共和国)
Zhōngguó Shèhuì Kēxuéyuàn(中国社会科学院)

yánjiūshēngyuàn(研究生院)
hóngshízìhuì(红十字会)
yúxīngcǎosù(鱼腥草素)
gǔshēngwùxuéjiā(古生物学家)

4.1.4 单音节词重叠，连写；双音节词重叠，分写。

rénrén(人人)　　niánnián(年年)
kànkan(看看)　　shuōshuo(说说)
dàdà(大大)　　hónghóng de(红红的)
gègè(个个)　　tiáotiáo(条条)

yánjiū yánjiū(研究研究)　　chángshì chángshì(尝试尝试)
xuěbái xuěbái(雪白雪白)　　tōnghóng tōnghóng(通红通红)

重叠并列即 AABB 式结构，当中加短横。

láilai-wǎngwǎng(来来往往)　　shuōshuo-xiàoxiào(说说笑笑)
qīngqīng-chǔchǔ(清清楚楚)　　wānwān-qūqū(弯弯曲曲)
jiājiā-hùhù(家家户户)　　qiānqiān-wànwàn(千千万万)

4.1.5 为了便于阅读和理解，在某些场合可以用短横。

huán-bǎo(环保——环境保护)　　gōng-guān(公关——公共关系)
bā-jiǔ tiān(八九天)　　shíqī-bā suì(十七八岁)
rén-jī duìhuà(人机对话)　　zhōng-xiǎoxué(中小学)
lù-hǎi-kōngjūn(陆海空军)　　biànzhèng-wéiwùzhǔyì(辩证唯物主义)

4.2 名词

4.2.1 名词与单音节前加成分(副、总、非、反、超、老、阿、可、无等)和单音节后加成分(子、儿、头、性、者、员、家、手、化、们等)，连写。

fùbùzhǎng(副部长)　　zǒnggōngchéngshī(总工程师)
fēijīnshǔ(非金属)　　fǎndàndào dǎodàn(反弹道导弹)
chāoshēngbō(超声波)　　fēiyèwù rényuán(非业务人员)

zhuōzi(桌子)　　mùtou(木头)

chéngwùyuán(乘务员)　　yìshùjiā(艺术家)

kēxuéxìng(科学性)　　xiàndàihuà(现代化)

háizimen(孩子们)　　tuōlājīshǒu(拖拉机手)

4.2.2　名词和后面的方位词，分写。

shān shàng(山上)　　shù xià(树下)

mén wài(门外)　　mén wàimian(门外面)

hé li(河里)　　hé lǐmian(河里面)

huǒchē shàngmian(火车上面)　　xuéxiào pángbiān(学校旁边)

Yǒngdìng Hé shàng(永定河上)　　Huáng Hé yǐnán(黄河以南)

但是，已经成词的，连写。例如："海外"不等于"海的外面"。

tiānshang(天上)　　dìxia(地下)

kōngzhōng(空中)　　hǎiwài(海外)

4.2.3　汉语人名按姓和名分写，姓和名的开头字母大写。笔名、别名等，按姓名写法处理。

Lǐ Huá(李华)　　Wáng Jiànguó(王建国)

Dōngfāng Shuò(东方朔)　　Zhūgě Kǒngmíng(诸葛孔明)

Lǔ Xùn(鲁迅)　　Méi Lánfāng(梅兰芳)

Zhāng Sān(张三)　　Wáng Mázi(王麻子)

姓名和职务、称呼等分开写；职务、称呼等开头小写。

Wáng bùzhǎng(王部长)　　Tián zhǔrèn(田主任)

Lǐ xiānsheng(李先生)　　Zhào tóngzhì(赵同志)

"老"、"小"、"大"、"阿"等称呼开头大写。

Xiǎo Liú(小刘)　　Lǎo Qián(老钱)

Dà Lǐ(大李)　　A Sān(阿三)

Wú Lǎo(吴老)

已经专名化的称呼，连写，开头大写。

Kǒngzǐ(孔子)　　Bāogōng(包公)

Xīshī(西施)　　Mèngchángjūn(孟尝君)

4.2.4　汉语地名按照中国地名委员会文件(84)中地字第17号《中国地名汉语拼音字母拼写规则(汉语地名部分)》的规定拼写。

汉语地名中的专名和通名分写，每一分写部分的第一个字母大写。

Běijīng Shì(北京市)　　Héběi Shěng(河北省)

Yālù Jiāng(鸭绿江)　　Tài Shān(泰山)

Dòngtíng Hú(洞庭湖)　　Táiwān Hǎixiá(台湾海峡)

专名和通名的附加成分，单音节的与其相关部分连写。

Xīliáo Hé(西辽河)　　Jǐngshān Hòujiē(景山后街)

Cháoyángménnèi Nánxiǎojiē(朝阳门内南小街)

自然村镇名称和其他不需区分专名和通名的地名，各音节连写。

Wángcūn(王村)　　Jiǔxiānqiáo(酒仙桥)

Zhōukǒudiàn(周口店)　　Sāntányìnyuè(三潭印月)

4.2.5　非汉语人名、地名本着"名从主人"的原则，按照罗马字母(拉丁字母)原文书写；非罗马字母文字的人名、地名，按照该文字的罗马字母转写法拼写。为了便于阅读，可以在原文后面注上汉字或汉字的拼音，在一定的场合也可以先用或仅用汉字的拼音。

Ulanhu(乌兰夫)　　Akutagawa Ryunosuke(芥川龙之介)
Ngapoi Ngawang Jigme(阿沛·阿旺晋美)　　Seypidin(赛福鼎)
Marx(马克思)　　Darwin(达尔文)
Neton(牛顿)　　Einstein(爱因斯坦)
Ürümqi(乌鲁木齐)　　Hohhot(呼和浩特)
Lhasa(拉萨)　　London(伦敦)
Paris(巴黎)　　Washington(华盛顿)
Tokyo(东京)

汉语化的音译名词，按汉字译音拼写。

Fēizhōu(非洲)　　Nánměi(南美)
Déguó(德国)　　Dōngnányà(东南亚)

4.3　动词

4.3.1　动词和“着”、“了”、“过”连写。

kànzhe(看着)　　jìnxíngzhe(进行着)
kànle(看了)　　jìnxíngle(进行了)
kànguo(看过)　　jìnxíngguo(进行过)

句末的“了”，分写。

Huǒchē dào le.(火车到了。)

4.3.2　动词和宾语，分写。

kàn xìn(看信)　　chī yú(吃鱼)
kāi wánxiào(开玩笑)　　jiāoliú jīngyàn(交流经验)

动宾式合成词中间插入其他成分的，分写。

jūle yī gè gōng(鞠了一个躬)　　lǐguo sān cì fà(理过三次发)

4.3.3　动词(或形容词)和补语，两者都是单音节的，连写；其余的情况，分写。

gǎohuài(搞坏)　　dǎsǐ(打死)
shútòu(熟透)　　jiànchéng(建成[楼房])
huàwéi(化为[蒸气])　　dàngzuò(当做[笑话])

zǒu jìnlai(走进来)　　zhěnglǐ hǎo(整理好)
jiànshè chéng(建设成[公园])　　gǎixiě wéi(改写为[剧本])

4.4　形容词

4.4.1　单音节形容词和重叠的前加成分或后加成分，连写。

mēngmēngliàng(蒙蒙亮)　　liàngtāngtāng(亮堂堂)

4.4.2　形容词和后面的“些”、“一些”、“点儿”、“一点儿”，分写。

dà xiē(大些)　　dà yīxiē(大一些)
kuài diǎnr(快点儿)　　kuài yīdiǎnr(快一点儿)

4.5　代词

4.5.1　表示复数的“们”和前面的代词，连写。

wǒmen(我们)　　tāmen(他们)

4.5.2　指示代词“这”、“那”，疑问代词“哪”和名词或量词，分写。

zhè rén(这人)　　nà cì huìyì(那次会议)
zhè zhī chuán(这只船)　　nǎ zhāng bàozhǐ(哪张报纸)

“这”、“那”、“哪”和“些”、“么”、“样”、“般”、“里”、“边”、“会儿”、“个”，连写。

zhèxiē(这些)　　zhème(这么)

nàyàng(那样)　　zhèbān(这般)

nàli(那里)　　nǎli(哪里)

zhèbiān(这边)　　zhèhuìr(这会儿)

zhège(这个)　　zhèmeyàng(这么样)

4.5.3 “各”、“每”、“某”、“本”、“该”、“我”、“你”等和后面的名词或量词，分写。

gè guó(各国)　　gè gè(各个)

gè rén(各人)　　gè xuékē(各学科)

měi nián(每年)　　měi cì(每次)

mǒu rén(某人)　　mǒu gōngchǎng(某工厂)

běn shì(本市)　　běn bùmén(本部门)

gāi kān(该刊)　　gāi gōngsī(该公司)

wǒ xiào(我校)　　nǐ dānwèi(你单位)

4.6 数词和量词

4.6.1 十一到九十九之间的整数，连写。

shíyī(十一)　　shíwǔ(十五)

sānshísān(三十三)　　jiǔshíjiǔ(九十九)

4.6.2 “百”、“千”、“万”、“亿”与前面的个位数，连写；“万”、“亿”与前面的十位以上的数，分写。

jiǔyì líng qīwàn èrqiān sānbǎi wǔshíliù(九亿零七万二千三百五十六)

liùshísān yì qīqiān èrbǎi liùshíbā wàn sìqiān líng jiǔshíwǔ(六十三亿七千二百六十八万四千零九十五)

4.6.3 表示序数的“第”与后面的数词中间，加短横。

dì-yī(第一)　　dì-shísān(第十三)

dì-èrshíbā(第二十八)　　dì-sānbǎi wǔshíliù(第三百五十六)

4.6.4 数词和量词，分写。

liǎng gè rén(两个人)　　yī dà wǎn fàn(一大碗饭)

liǎng jiān bàn wūzi(两间半屋子)　　wǔshísān rénci(五十三人次)

表示约数的“多”、“来”、“几”和数词、量词分写。

yìbǎi duō gè(一百多个)　　shí lái wàn rén(十来万人)

jǐ jiā rén(几家人)　　jǐ tiān gōngfu(几天工夫)

“十几”、“几十”连写。

shíjǐ gè rén(十几个人)　　jǐshí gēn gāngguǎn(几十根钢管)

4.7 虚词

虚词与其他语词分写。

4.7.1 副词

hěn hǎo(很好)　　dōu lái(都来)

gèng měi(更美)　　zuì dà(最大)

bù lái(不来)

yīng bù yīnggāi(应不应该)　　gānggāng zǒu(刚刚走)

fēicháng kuài(非常快)　　shífēn gǎndòng(十分感动)

4.7.2 介词

zài qiánmiàn(在前面)　　xiàng dōngbiān qù(向东边去)

wèi rénmín fúwù(为人民服务)　　cóng zuótiān qǐ(从昨天起)
shēng yú 1940 nián(生于1940年)　　guānyú zhège wèntí(关于这个问题)

4.7.3 连词

gōngrén hé nóngmín(工人和农民)　　bùdàn kuài érqiě hǎo(不但快而且好)
guāngróng ér jiānjù(光荣而艰巨)　　Nǐ lái háishi bù lái?(你来还是不来?)

4.7.4 结构助词"的"、"地"、"得"、"之"

dàdì de nǚ'ér(大地的女儿)
Zhè shì wǒ de shū.(这是我的书。)
Wǒmen guòzhe xìngfú de shēnghuó.(我们过着幸福的生活。)
Shāngdiàn li bǎimǎnle chī de,chuān de,yòng de.(商店里摆满了吃的、穿的、用的。)
mài qīngcài luóbo de(卖青菜萝卜的)

Tā zài dàjiē shàng mànman de zǒu.(他在大街上慢慢地走。)
Tǎnbái de gàosu nǐ ba.(坦白地告诉你吧。)
Tā yī bù yī gè jiǎoyìnr de gōngzuòzhe.(他一步一个脚印儿地工作着。)

dǎsǎo de gānjìng(打扫得干净)　　xiě de bù hǎo(写得不好)
hóng de hěn(红得很)　　lěng de fādǒu(冷得发抖)

shàonián zhī jiā(少年之家)
zuì fādá de guójiā zhī yī(最发达的国家之一)

注:"的"、"地"、"得"在技术处理上,根据需要可以分别写作"d"、"di"、"de"。

4.7.5 语气助词

Nǐ zhīdao ma?(你知道吗?)
Zěnme hái bù lái a?(怎么还不来啊?)
Kuài qù ba!(快去吧!)
Tā shì bù huì lái de.(他是不会来的。)

4.7.6 叹词

A!Zhēn měi!(啊!真美!)
Ng,nǐ shuō shénme?(嗯,你说什么?)
Hm,zǒuzhe qiáo ba!(哼,走着瞧吧!)

4.7.7 拟声词

pa!(啪!)　　huahua(哗哗)
jiji-zhazha(叽叽喳喳)　　"honglong" yī shēng("轰隆"一声)
Dà gōngjī wo—wo—tí.(大公鸡喔喔啼。)
"Du—"qìdí xiǎng le.("嘟——"汽笛响了。)

4.8 成语

4.8.1 四言成语可以分为两个双音节来念的,中间加短横。

céngchū-bùqióng(层出不穷)　　fēngpíng-làngjìng(风平浪静)
àizēng-fēnmíng(爱憎分明)　　shuǐdào-qúchéng(水到渠成)
yángyáng-dàguān(洋洋大观)　　píngfēn-qiūsè(平分秋色)
guāngmíng-lěiluò(光明磊落)　　diānsān-dǎosì(颠三倒四)

4.8.2 不能按两段来念的四言成语、熟语等,全部连写。

bùyìlèhū(不亦乐乎)　　zǒng'éryánzhī(总而言之)
àimònéngzhù(爱莫能助)　　yīyīdàishuǐ(一衣带水)
húlihútu(糊里糊涂)　　hēibuliūqiū(黑不溜秋)
diào'erlángdāng(吊儿郎当)

4.9　大写

4.9.1　句子开头的字母和诗歌每行开头的字母大写。(举例略)

4.9.2　专有名词的第一个字母大写。

Běijīng(北京)　Chángchéng(长城)　Qīngmíng(清明)

由几个词组成的专有名词,每个词的第一个字母大写。

Guójì Shūdiàn(国际书店)　　Hépíng Bīnguǎn(和平宾馆)
Guāngmíng Rìbào(光明日报)

4.9.3　专有名词和普通名词连写在一起的,第一个字母要大写。

Zhōngguórén(中国人)　　Míngshǐ(明史)
Guǎngdōnghuà(广东话)

已经转化为普通名词的,第一个字母小写。

guǎnggān(广柑)　　zhōngshānfú(中山服)
chuānxiōng(川芎)　　zàngqīngguǒ(藏青果)

4.10　移行

4.10.1　移行要按音节分开,在没有写完的地方加上短横。

……guāng-
míng(光明)

不能移作"gu-āngmíng"。

4.11　标调

4.11.1　声调一律标原调,不标变调。

yī jià(一架)	yī tiān(一天)	yī tóu(一头)
yī wǎn(一碗)	qī wàn(七万)	qī běn(七本)
bā gè(八个)	qī shàng-bāxià(七上八下)	
bù qù(不去)	bù duì(不对)	bùzhìyú(不至于)

但是在语音教学时可以根据需要按变调标写。

注:除了《汉语拼音方案》规定的符号标调法以外,在技术处理上,也可根据需要采用数字或字母作为临时变通标调法。

附加说明:

本标准由国家教育委员会、国家语言文字工作委员会提出。

本标准由汉语拼音正词法委员会负责起草。

本标准主要起草人尹斌庸、李乐毅、金惠淑。

2. Základní pravopisná pravidla čínské fonetické abecedy, norma GB/T

ICS 01.140.10
A 14

中华人民共和国国家标准

GB/T 16159—2012
代替 GB/T 16159—1996

汉语拼音正词法基本规则

Basic rules of the Chinese phonetic alphabet orthography

2012-06-29 发布　　2012-10-01 实施

中华人民共和国国家质量监督检验检疫总局
中国国家标准化管理委员会　发布

数码防伪

16159–2012

目　次

前　言

本标准按照 GB/T 1.1—2009 给出的规则起草。

本标准代替 GB/T 16159—1996《汉语拼音正词法基本规则》。

本标准与 GB/T 16159—1996 相比，主要变化如下：

——将原标准中正词法的具体规定、用法调整为分词连写、人名地名拼写、大写、缩写、标调、移行、标点符号使用等 7 个部分的基本规则。其中，把原先按词类分节的部分归到分词连写规则之下，并增加了“缩写规则”和“标点符号使用规则”。

——取消原标准中与名词、动词、形容词、代词、数词和量词并列的“虚词”一节，把虚词词类提升，与实词词类并列，以贯彻按词类分节的原则。

——修改了原标准中关于非汉语人名、地名的汉语拼音拼写规则。

——参照 ISO 7098《中文罗马字母拼写法》的规定，补充了“汉字数字用汉语拼音拼写，阿拉伯数字则仍保留阿拉伯数字写法”的规定。

——增加了在某些场合，专有名词的所有字母可全部大写，也可不标声调的规定。

——增加了变通规则，以照顾某些领域的特殊需要。

本标准由教育部语言文字信息管理司提出并归口。

本标准主要起草单位：中国社会科学院语言研究所、教育部语言文字应用研究所。

本标准主要起草人：董琨、李志江、金惠淑、史定国、王楠、杜翔。

汉语拼音正词法基本规则

1 范围

本标准规定了用《汉语拼音方案》拼写现代汉语的规则。内容包括分词连写规则、人名地名拼写规则、大写规则、标调规则、移行规则、标点符号使用规则等。为了适应特殊的需要，同时规定了一些变通规则。

本标准适用于文化教育、编辑出版、中文信息处理及其他方面的汉语拼音拼写。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本(包括所有的修改单)适用于本文件。

GB/T 15834 标点符号用法

GB/T 28039 中国人名汉语拼音字母拼写规则

《汉语拼音方案》(1958年2月11日第一届全国人民代表大会第五次会议批准)

《中国地名汉语拼音字母拼写规则(汉语地名部分)》(1984年12月25日中国地名委员会、中国文字改革委员会、国家测绘局发布)

3 术语和定义

下列术语和定义适用于本文件。

3.1

词 word

语言里最小的、可以独立运用的单位。

3.2

汉语拼音方案 scheme for the Chinese phonetic alphabet

给汉字注音和拼写普通话语音的方案，1958年2月11日第一届全国人民代表大会第五次会议批准。方案采用拉丁字母，并用附加符号表示声调，是帮助学习汉字和推广普通话的工具。

3.3

汉语拼音正词法 the Chinese phonetic alphabet orthography

汉语拼音的拼写规范及其书写格式的准则。

4 制定原则

4.1 本标准是在《汉语拼音方案》确定的音节拼写规则的基础上进一步规定的词的拼写规则。

4.2 以词为拼写单位，适当考虑语音、语义等因素，并兼顾词的拼写长度。

4.3 按语法词类分节规定分词连写规则。

5 总则

5.1 拼写普通话基本上以词为书写单位。例如：

rén（人）　　pǎo（跑）
hǎo（好）　　nǐ（你）
sān（三）　　gè（个）
hěn（很）　　bǎ（把）
hé（和）　　de（的）
ā（啊）　　pēng（砰）
fúróng（芙蓉）　　qiǎokèlì（巧克力）
māma（妈妈）　　péngyou（朋友）
yuèdú（阅读）　　wǎnhuì（晚会）
zhòngshì（重视）　　dìzhèn（地震）
niánqīng（年轻）　　qiānmíng（签名）
shìwēi（示威）　　niǔzhuǎn（扭转）
chuánzhī（船只）　　dànshì（但是）
fēicháng（非常）　　dīngdōng（叮咚）
āiyā（哎呀）　　diànshìjī（电视机）
túshūguǎn（图书馆）

5.2 表示一个整体概念的双音节和三音节结构，连写。例如：

quánguó（全国）　　zǒulái（走来）
dǎnxiǎo（胆小）　　huánbǎo（环保）
gōngguān（公关）　　chángyòngcí（常用词）
àiniǎozhōu（爱鸟周）　　yǎnzhōngdīng（眼中钉）
èzuòjù（恶作剧）　　pòtiānhuāng（破天荒）
yīdāoqiē（一刀切）　　duìbuqǐ（对不起）
chīdexiāo（吃得消）

5.3 四音节及四音节以上表示一个整体概念的名称，按词或语节（词语内部由语音停顿而划分成的片段）分写，不能按词或语节划分的，全都连写。例如：

wúfèng gāngguǎn（无缝钢管）　　huánjìng bǎohù guīhuà（环境保护规划）
jīngtǐguǎn gōnglǜ fàngdàqì（晶体管功率放大器）
Zhōnghuá Rénmín Gònghéguó（中华人民共和国）
Zhōngguó Shèhuì Kēxuéyuàn（中国社会科学院）
yánjiūshēngyuàn（研究生院）
hóngshízìhuì（红十字会）　　yúxīngcǎosù（鱼腥草素）
gāoměngsuānjiǎ（高锰酸钾）　　gǔshēngwùxuéjiā（古生物学家）

5.4 单音节词重叠，连写；双音节词重叠，分写。例如：

rénrén（人人）　　niánnián（年年）
kànkan（看看）　　shuōshuo（说说）
dàdà（大大）　　hónghóng de（红红的）

gègè（个个）　　tiáotiáo（条条）
yánjiū yánjiū（研究研究）　　shāngliang shāngliang（商量商量）
xuěbái xuěbái（雪白雪白）　　tōnghóng tōnghóng（通红通红）

重叠并列即AABB式结构，连写。例如：

láilaiwǎngwǎng（来来往往）　　shuōshuōxiàoxiào（说说笑笑）
qīngqīngchǔchǔ（清清楚楚）　　wānwānqūqū（弯弯曲曲）
fāngfāngmiànmiàn（方方面面）　　qiānqiānwànwàn（千千万万）

5.5 单音节前附成分（副、总、非、反、超、老、阿、可、无、半等）或单音节后附成分（子、儿、头、性、者、员、家、手、化、们等）与其他词语，连写。例如：

fùbùzhǎng（副部长）　　zǒnggōngchéngshī（总工程师）
fùzǒnggōngchéngshī（副总工程师）　　fēijīnshǔ（非金属）
fēiyèwù rényuán（非业务人员）　　fǎndàndào dǎodàn（反弹道导弹）
chāoshēngbō（超声波）　　lǎohǔ（老虎）
āyí（阿姨）　　kěnì fǎnyìng（可逆反应）
wútiáojiàn（无条件）　　bàndǎotǐ（半导体）
zhuōzi（桌子）　　jīnr（今儿）
quántou（拳头）　　kēxuéxìng（科学性）
shǒugōngyèzhě（手工业者）　　chéngwùyuán（乘务员）
yìshùjiā（艺术家）　　tuōlājīshǒu（拖拉机手）
xiàndàihuà（现代化）　　háizimen（孩子们）

5.6 为了便于阅读和理解，某些并列的词、语素之间或某些缩略语当中可用连接号。例如：

bā-jiǔ tiān（八九天）　　shíqī-bā suì（十七八岁）
rén-jī duìhuà（人机对话）　　zhōng-xiǎoxué（中小学）
lù-hǎi-kōngjūn（陆海空军）　　biànzhèng-wéiwù zhǔyì（辩证唯物主义）
Cháng-Sānjiǎo（长三角[长江三角洲]）　　Hù-Níng-Háng Dìqū（沪宁杭地区）
Zhè-Gàn Xiàn（浙赣线）　　Jīng-Zàng Gāosù Gōnglù（京藏高速公路）

6 基本规则

6.1 分词连写规则

6.1.1 名词

6.1.1.1 名词与后面的方位词，分写。例如：

shān shàng（山上）　　shù xià（树下）
mén wài（门外）　　mén wàimian（门外面）
hé li（河里）　　hé lǐmian（河里面）
huǒchē shàngmian（火车上面）　　xuéxiào pángbiān（学校旁边）
Yǒngdìng Hé shàng（永定河上）　　Huáng Hé yǐnán（黄河以南）

6.1.1.2 名词与后面的方位词已经成词的，连写。例如：

tiānshang（天上）　　dìxia（地下）
kōngzhōng（空中）　　hǎiwài（海外）

6.1.2 动词

6.1.2.1 动词与后面的动态助词“着”、“了”、“过”，连写。例如：

kànzhe（看着）　　tǎolùn bìng tōngguòle（讨论并通过了）

jìnxíngguo（进行过）

6.1.2.2 句末的“了”兼做语气助词，分写。例如：

Zhè běn shū wǒ kàn le.（这本书我看了。）

6.1.2.3 动词与所带的宾语，分写。例如：

kàn xìn（看信）　　chī yú（吃鱼）

kāi wánxiào（开玩笑）　　jiāoliú jīngyàn（交流经验）

动宾式合成词中间插入其他成分的，分写。

jūle yī gè gōng（鞠了一个躬）　　lǐguo sān cì fà（理过三次发）

6.1.2.4 动词（或形容词）与后面的补语，两者都是单音节的，连写；其余情况，分写。例如：

gǎohuài（搞坏）　　dǎsǐ（打死）

shútòu（熟透）　　jiànchéng（建成[楼房]）

huàwéi（化为[蒸汽]）　　dàngzuò（当作[笑话]）

zǒu jìnlái（走进来）　　zhěnglǐ hǎo（整理好）

jiànshè chéng（建设成[公园]）　　gǎixiě wéi（改写为[剧本]）

6.1.3 形容词

6.1.3.1 单音节形容词与用来表示形容词生动形式的前附成分或后附成分，连写。例如：

mēngmēngliàng（蒙蒙亮）　　liàngtángtáng（亮堂堂）

hēigulōngdōng（黑咕隆咚）

6.1.3.2 形容词与后面的“些”“一些”“点儿”“一点儿”，分写。例如：

dà xiē（大些）　　dà yīxiē（大一些）

kuài diǎnr（快点儿）　　kuài yīdiǎnr（快一点儿）

6.1.4 代词

6.1.4.1 人称代词、疑问代词与其他词语，分写。例如：

Wǒ ài Zhōngguó.（我爱中国。）　　Tāmen huílái le.（他们回来了。）

Shéi shuō de?（谁说的?）　　Qù nǎlǐ?（去哪里?）

6.1.4.2 指示代词“这”、“那”，疑问代词“哪”与后面的名词或量词，分写。例如：

zhè rén（这人）　　nà cì huìyì（那次会议）

zhè zhī chuán（这只船）　　nǎ zhāng bàozhǐ（哪张报纸）

指示代词“这”、“那”，疑问代词“哪”与后面的“点儿”、“般”、“边”、“时”、“会儿”，连写。例如：

zhèdiǎnr（这点儿）　　zhèbān（这般）

zhèbiān（这边）　　nàshí（那时）

nàhuìr（那会儿）

6.1.4.3 “各”、“每”、“某”、“本”、“该”、“我”、“你”等与后面的名词或量词，分写。例如：

gè guó（各国）　　gè rén（各人）

gè xuékē（各学科）　　měi nián（每年）

měi cì（每次）　　mǒu rén（某人）
mǒu gōngchǎng（某工厂）　　běn shì（本市）
běn bùmén（本部门）　　gāi kān（该刊）
gāi gōngsī（该公司）　　wǒ xiào（我校）
nǐ dānwèi（你单位）

6.1.5　数词和量词

6.1.5.1　汉字数字用汉语拼音拼写，阿拉伯数字则仍保留阿拉伯数字写法。例如：

èr líng líng bā nián（二〇〇八年）　　èr fēn zhī yī（二分之一）
wǔ yòu sì fēn zhī sān（五又四分之三）　　sān diǎn yī sì yī liù（三点一四一六）
líng diǎn liù yī bā（零点六一八）　　635 fēnjī（635 分机）

6.1.5.2　十一到九十九之间的整数，连写。例如：

shíyī（十一）　　shíwǔ（十五）
sānshísān（三十三）　　jiǔshíjiǔ（九十九）

6.1.5.3　“百”“千”“万”“亿”与前面的个位数，连写；“万”“亿”与前面的十位以上的数，分写，当前面的数词为“十”时，也可连写。例如：

shí yì líng qīwàn èrqiān sānbǎi wǔshíliù/shíyì líng qīwàn èrqiān sānbǎi wǔshíliù
（十亿零七万二千三百五十六）
liùshísān yì qīqiān èrbǎi liùshíbā wàn sìqiān líng jiǔshíwǔ
（六十三亿七千二百六十八万四千零九十五）

6.1.5.4　数词与前面表示序数的“第”中间，加连接号。例如：

dì-yī（第一）　　dì-shísān（第十三）
dì-èrshíbā（第二十八）　　dì-sānbǎi wǔshíliù（第三百五十六）

数词（限于“一”至“十”）与前面表示序数的“初”，连写。例如：

chūyī（初一）　　chūshí（初十）

6.1.5.5　代表月日的数词，中间加连接号。例如：

wǔ-sì（五四）　　yīèr-jiǔ（一二·九）

6.1.5.6　数词与量词，分写。例如：

liǎng gè rén（两个人）　　yī dà wǎn fàn（一大碗饭）
liǎng jiān bàn wūzi（两间半屋子）　　kàn liǎng biàn（看两遍）

数词、量词与表示约数的“多”、“来”、“几”，分写。

yībǎi duō gè（一百多个）　　shí lái wàn rén（十来万人）
jǐ jiā rén（几家人）　　jǐ tiān gōngfu（几天工夫）

“十几”、“几十”连写。例如：

shíjǐ gè rén（十几个人）　　jǐshí gēn gāngguǎn（几十根钢管）

两个邻近的数字或表位数的单位并列表示约数，中间加连接号。例如：

sān-wǔ tiān（三五天）　　qī-bā gè（七八个）
yì-wàn nián（亿万年）　　qiān-bǎi cì（千百次）

复合量词内各并列成分连写。例如：

réncì（人次）　　qiānwǎxiǎoshí（千瓦小时）
dūngōnglǐ（吨公里）　　qiānkèmǐměimiǎo（千克·米/秒）

6.1.6 副词

副词与后面的词语，分写。例如：

hěn hǎo（很好）
gèng měi（更美）
bù lái（不来）
gānggāng zǒu（刚刚走）
shífēn gǎndòng（十分感动）
dōu lái（都来）
zuì dà（最大）
bù hěn hǎo（不很好）
fēicháng kuài（非常快）

6.1.7 介词

介词与后面的其他词语，分写。例如：

zài qiánmiàn zǒu（在前面走）
wèi rénmín fúwù（为人民服务）
bèi xuǎnwéi dàibiǎo（被选为代表）
guānyú zhège wèntí（关于这个问题）
xiàng dōngbian qù（向东边去）
cóng zuótiān qǐ（从昨天起）
shēng yú 1940 nián（生于 1940 年）
cháozhe xiàbian kàn（朝着下边看）

6.1.8 连词

连词与其他词语，分写。例如：

gōngrén hé nóngmín（工人和农民）
guāngróng ér jiānjù（光荣而艰巨）
tóngyì bìng yōnghù（同意并拥护）
bùdàn kuài érqiě hǎo（不但快而且好）
Nǐ lái háishi bù lái?（你来还是不来?）
Rúguǒ xià dàyǔ,bǐsài jiù tuīchí.（如果下大雨，比赛就推迟。）

6.1.9 助词

6.1.9.1 结构助词“的”、“地”、“得”、“之”、“所”等与其他词语，分写。其中，“的”、“地”、“得”前面的词是单音节的，也可连写。例如：

dàdì de nǚ'ér（大地的女儿）
Zhè shì wǒ de shū./Zhè shì wǒde shū.（这是我的书。）
Wǒmen guòzhe xìngfú de shēnghuó.（我们过着幸福的生活。）
Shāngdiàn li bǎimǎnle chī de, chuān de, yòng de./Shāngdiàn li bǎimǎnle chīde, chuānde, yòngde.（商店里摆满了吃的、穿的、用的。）
mài qīngcài luóbo de（卖青菜萝卜的）
Tā zài dàjiē shang mànman de zǒu.（他在大街上慢慢地走。）
Tǎnbái de gàosu nǐ ba.（坦白地告诉你吧。）
Tā yī bù yī gè jiǎoyìnr de gōngzuòzhe.（他一步一个脚印儿地工作着。）
dǎsǎo de gānjìng（打扫得干净）
hóng de hěn /hóngde hěn（红得很）
shàonián zhī jiā（少年之家）
jù wǒ suǒ zhī（据我所知）
bèi yīngxióng de shìjì suǒ gǎndòng（被英雄的事迹所感动）
xiě de bù hǎo /xiěde bù hǎo（写得不好）
lěng de fādǒu/lěngde fādǒu（冷得发抖）
zuì fādá de guójiā zhī yī（最发达的国家之一）

6.1.9.2 语气助词与其他词语，分写。例如：

Nǐ zhīdào ma?（你知道吗?）　　Zěnme hái bù lái a?（怎么还不来啊?）
Kuài qù ba!（快去吧!）　　Tā yīdìng huì lái de.（他一定会来的。）
Huǒchē dào le.（火车到了。）
Tā xīnli míngbai，zhǐshì bù shuō bàle.（他心里明白，只是不说罢了。）

6.1.9.3 动态助词

动态助词主要有“着”、“了”、“过”。见 6.1.2.1 的规定。

6.1.10 叹词

叹词通常独立于句法结构之外，与其他词语分写。例如：

À! Zhēn měi!（啊! 真美!）　　Ńg，nǐ shuō shénme?（嗯，你说什么?）
Hng，zǒuzhe qiáo ba!（哼，走着瞧吧!）　　Tīng míngbai le ma? Wèi!（听明白了吗? 喂!）
Āiyā，wǒ zěnme bù zhīdào ne!（哎呀，我怎么不知道呢!）

6.1.11 拟声词

拟声词与其他词语，分写。例如：

“hōnglōng” yī shēng（“轰隆”一声）　　chánchán liúshuǐ（潺潺流水）
mó dāo huòhuò（磨刀霍霍）　　jījīzhāzhā jiào gè bù tíng（叽叽喳喳叫个不停）
Dà gōngjī wōwō tí.（大公鸡喔喔啼。）　　“Dū——”，qìdí xiǎng le.（“嘟——”，汽笛响了。）
Xiǎoxī huāhuā de liútǎng.（小溪哗哗地流淌。）

6.1.12 成语和其他熟语

6.1.12.1 成语通常作为一个语言单位使用，以四字文言语句为主。结构上可以分为两个双音节的，中间加连接号。例如：

fēngpíng-làngjìng（风平浪静）　　àizēng-fēnmíng（爱憎分明）
shuǐdào-qúchéng（水到渠成）　　yángyáng-dàguān（洋洋大观）
píngfēn-qiūsè（平分秋色）　　guāngmíng-lěiluò（光明磊落）
diānsān-dǎosì（颠三倒四）

结构上不能分为两个双音节的，全部连写。例如：

céngchūbùqióng（层出不穷）　　bùyìlèhū（不亦乐乎）
zǒng'éryánzhī（总而言之）　　àimònéngzhù（爱莫能助）
yīyīdàishuǐ（一衣带水）

6.1.12.2 非四字成语和其他熟语内部按词分写。例如：

bēi hēiguō（背黑锅）　　yī bíkǒng chū qì（一鼻孔出气儿）
bā gānzi dǎ bù zháo（八竿子打不着）
zhǐ xǔ zhōuguān fàng huǒ，bù xǔ bǎixìng diǎn dēng（只许州官放火，不许百姓点灯）
xiǎocōng bàn dòufu——yīqīng-èrbái（小葱拌豆腐——一清二白）

6.2 人名地名拼写规则

6.2.1 人名拼写

6.2.1.1 汉语人名中的姓和名分写，姓在前，名在后。复姓连写。双姓中间加连接号。姓和名的首字母分别大写，双姓两个字首字母都大写。笔名、别名等，按姓名写法处理。例如：

Lǐ Huá（李华）
Dōngfāng Shuò（东方朔）
Zhāng-Wáng Shūfāng（张王淑芳）
Méi Lánfāng（梅兰芳）
Wáng Mázi（王麻子）
Wáng Jiànguó（王建国）
Zhūgě Kǒngmíng（诸葛孔明）
Lǔ Xùn（鲁迅）
Zhāng Sān（张三）

6.2.1.2 人名与职务、称呼等，分写；职务、称呼等首字母小写。例如：

Wáng bùzhǎng（王部长）
Wú kuàijì（吴会计）
Zhào tóngzhì（赵同志）
Dīng xiōng（丁兄）
Zhāng jūn（张君）
Wáng shì（王氏）
Guóqiáng tóngzhì（国强同志）
Tián zhǔrèn（田主任）
Lǐ xiānsheng（李先生）
Liú lǎoshī（刘老师）
Zhāng mā（张妈）
Wú lǎo（吴老）
Sūn mǒu（孙某）
Huìfāng āyí（慧芳阿姨）

6.2.1.3 “老”、“小”、“大”、“阿”等与后面的姓、名、排行，分写，分写部分的首字母分别大写。例如：

Xiǎo Liú（小刘）
Lǎo Zhāngtour（老张头儿）
Ā Sān（阿三）
Lǎo Qián（老钱）
Dà Lǐ（大李）

6.2.1.4 已经专名化的称呼，连写，开头大写。例如：

Kǒngzǐ（孔子）
Xīshī（西施）
Bāogōng（包公）
Mèngchángjūn（孟尝君）

6.2.2 地名拼写

6.2.2.1 汉语地名中的专名和通名，分写，每一分写部分的首字母大写。例如：

Běijīng Shì（北京市）
Yālù Jiāng（鸭绿江）
Dòngtíng Hú（洞庭湖）
Héběi Shěng（河北省）
Tài Shān（泰山）
Táiwān Hǎixiá（台湾海峡）

6.2.2.2 专名与通名的附加成分，如是单音节的，与其相关部分连写。例如：

Xīliáo Hé（西辽河）
Cháoyángménnèi Nánxiǎojiē（朝阳门内南小街）
Jǐngshān Hòujiē（景山后街）
Dōngsì Shítiáo（东四十条）

6.2.2.3 已专名化的地名不再区分专名和通名，各音节连写。例如：

Hēilóngjiāng（黑龙江[省]）
Jiǔxiānqiáo（酒仙桥[医院]）
Wángcūn（王村[镇]）

不需区分专名和通名的地名，各音节连写。例如：

Zhōukǒudiàn（周口店）
Sāntányìnyuè（三潭印月）

6.2.3 非汉语人名、地名的汉字名称，用汉语拼音拼写。例如：

Wūlánfū（乌兰夫，Ulanhu）
Jièchuān Lóngzhījiè（芥川龙之介，Akutagawa Ryunosuke）
Āpèi Āwàngjìnměi（阿沛·阿旺晋美，Ngapoi Ngawang Jigme）
Mǎkèsī（马克思，Marx）
Lúndūn（伦敦，London）
Wūlǔmùqí（乌鲁木齐，Ürümqi）
Dōngjīng（东京，Tokyo）

6.2.4 人名、地名拼写的详细规则，遵循 GB/T 28039《中国人名汉语拼音字母拼写规则》《中国地名汉语拼音字母拼写规则(汉语地名部分)》。

6.3 大写规则

6.3.1 句子开头的字母大写。例如：

Chūntiān lái le.(春天来了。)

Wǒ ài wǒ de jiāxiāng.(我爱我的家乡。)

诗歌每行开头的字母大写。例如：

《Yǒude Rén》(《有的人》)

Zāng Kèjiā(臧克家)

Yǒude rén huózhe,(有的人活着，)

Tā yǐjīng sǐ le;(他已经死了；)

Yǒude rén sǐ le,(有的人死了，)

Tā hái huózhe.(他还活着。)

6.3.2 专有名词的首字母大写。例如：

Běijīng(北京)　　Chángchéng(长城)

Qīngmíng(清明)　　Jǐngpōzú(景颇族)

Fēilǜbīn(菲律宾)

由几个词组成的专有名词，每个词的首字母大写。例如：

Guójì Shūdiàn(国际书店)　　Hépíng Bīnguǎn(和平宾馆)

Guāngmíng Rìbào(光明日报)

Guójiā Yǔyán Wénzì Gōngzuò Wěiyuánhuì(国家语言文字工作委员会)

在某些场合，专有名词的所有字母可全部大写。例如：

XIÀNDÀI HÀNYǓ CÍDIǍN(现代汉语词典)　　BĚIJĪNG(北京)

LǏ HUÁ(李华)　　DŌNGFĀNG SHUÒ(东方朔)

6.3.3 专有名词成分与普通名词成分连写在一起，是专有名词或视为专有名词的，首字母大写。例如：

Míngshǐ(明史)　　Hànyǔ(汉语)

Yuèyǔ(粤语)　　Guǎngdōnghuà(广东话)

Fójiào(佛教)　　Tángcháo(唐朝)

专有名词成分与普通名词成分连写在一起，是一般语词或视为一般语词的，首字母小写。例如：

guǎnggān(广柑)　　jīngjù(京剧)

ējiāo(阿胶)　　zhōngshānfú(中山服)

chuānxiōng(川芎)　　zàngqīngguǒ(藏青果)

zhāoqín-mùchǔ(朝秦暮楚)　　qiánlǘzhījì(黔驴之技)

6.4 缩写规则

6.4.1 连写的拼写单位(多音节词或连写的表示一个整体概念的结构)，缩写时取每个汉字拼音的首字母，大写并连写。例如：

Běijīng(缩写：BJ)(北京)　　ruǎnwò(缩写：RW)(软卧)

6.4.2 分写的拼写单位(按词或语节分写的表示一个整体概念的结构),缩写时以词或语节为单位取首字母,大写并连写。例如:

guójiā biāozhǔn (缩写:GB) (国家标准)

hànyǔ shuǐpíng kǎoshì (缩写:HSK) (汉语水平考试)

pǔtōnghuà shuǐpíng cèshì (缩写:PSC) (普通话水平测试)

6.4.3 为了给汉语拼音的缩写形式做出标记,可在每个大写字母后面加小圆点。例如:

Běijīng (北京)也可缩写:B. J.　　guójiā biāozhǔn (国家标准)也可缩写:G. B.

6.4.4 汉语人名的缩写,姓全写,首字母大写或每个字母大写;名取每个汉字拼音的首字母,大写,后面加小圆点。例如:

Lǐ Huá (缩写:Lǐ H. 或 LǏ H.) (李华)

Wáng Jiànguó (缩写:Wáng J. G. 或 WÁNG J. G.) (王建国)

Dōngfāng Shuò (缩写:Dōngfāng S. 或 DŌNGFĀNG S.) (东方朔)

Zhūgě Kǒngmíng (缩写:Zhūgě K. M. 或 ZHŪGĚ K. M.) (诸葛孔明)

6.5 标调规则

6.5.1 声调符号标在一个音节的主要元音(韵腹)上。韵母 iu,ui,声调符号标在后面的字母上面。在 i 上标声调符号,应省去 i 上的小点。例如:

āyí (阿姨)　　cèlüè (策略)

dàibiǎo (代表)　　guāguǒ (瓜果)

huáishù (槐树)　　kǎolǜ (考虑)

liúshuǐ (流水)　　xīnxiān (新鲜)

轻声音节不标声调。例如:

zhuāngjia (庄稼)　　qīngchu (清楚)

kàndeqǐ (看得起)

6.5.2 “一”、“不”一般标原调,不标变调。例如:

yī jià (一架)　　yī tiān (一天)

yī tóu (一头)　　yī wǎn (一碗)

bù qù (不去)　　bù duì (不对)

bùzhìyú (不至于)

在语言教学等方面,可根据需要按变调标写。例如:

yī tiān (一天)可标为 yì tiān,bù duì (不对)可标为 bú duì。

6.5.3 ABB、AABB 形式的词语,BB 一般标原调,不标变调。例如:

lǜyóuyóu (绿油油)　　chéndiàndiàn (沉甸甸)

hēidòngdòng (黑洞洞)　　piàopiàoliàngliàng (漂漂亮亮)

有些词语的 BB 在语言实际中只读变调,则标变调。例如:

hóngtōngtōng (红彤彤)　　xiāngpēnpēn (香喷喷)

huángdēngdēng (黄澄澄)

6.5.4 在某些场合,专有名词的拼写,也可不标声调。例如:

Li Hua (缩写:Li H. 或 LI H.) (李华)　　Beijing (北京)

RENMIN RIBAO (人民日报)　　WANGFUJING DAJIE (王府井大街)

6.5.5 除了《汉语拼音方案》规定的符号标调法以外，在技术处理上，也可采用数字、字母等标明声调，如采用阿拉伯数字 1、2、3、4、0 分别表示汉语四声和轻声。

6.6 移行规则

6.6.1 移行要按音节分开，在没有写完的地方加连接号。音节内部不可拆分。例如：

guāngmíng（光明）移作“……guāng-
míng”（光明）
不能移作“……gu-
āngmíng”（光明）。

缩写词（如 GB，HSK，汉语人名的缩写部分）不可移行。

Wáng J. G.（王建国）移作“……Wáng
J. G.”（王建国）
不能移作“……Wáng J.-
G.”（王建国）。

6.6.2 音节前有隔音符号，移行时，去掉隔音符号，加连接号。例如：

Xī'ān（西安）移作“……Xī-
ān”（西安）
不能移作“……Xī'-
ān”（西安）。

6.6.3 在有连接号处移行时，末尾保留连接号，下行开头补加连接号。例如：

chēshuǐ-mǎlóng（车水马龙）移作“……chēshuǐ-
-mǎlóng”（车水马龙）

6.7 标点符号使用规则

汉语拼音拼写时，句号使用小圆点“.”，连接号用半字线“-”，省略号也可使用 3 个小圆点“…”，顿号也可用逗号“,”代替，其他标点符号遵循 GB/T 15834 的规定。

7 变通规则

7.1 根据识字需要（如小学低年级和幼儿汉语识字读物），可按字注音。

7.2 辞书注音需要显示成语及其他词语内部结构时，可按词或语素分写。例如：

chīrén shuō mèng（痴人说梦）	wèi yǔ chóumóu（未雨绸缪）
shǒu kǒu rú píng（守口如瓶）	Hēng-Hā èr jiàng（哼哈二将）
Xī Liáo Hé（西辽河）	Nán-Běi Cháo（南北朝）

7.3 辞书注音为了提示轻声音节，音节前可标中圆点。例如：

zhuāng · jia（庄稼）	qīng · chu（清楚）
kàn · deqǐ（看得起）	

如是轻重两读，音节上仍标声调。例如：

hóu · lóng（喉咙）	zhī · dào（知道）
tǔ · xīngqì（土腥气）	

7.4 在中文信息处理方面，表示一个整体概念的多音节结构，可全部连写。例如：

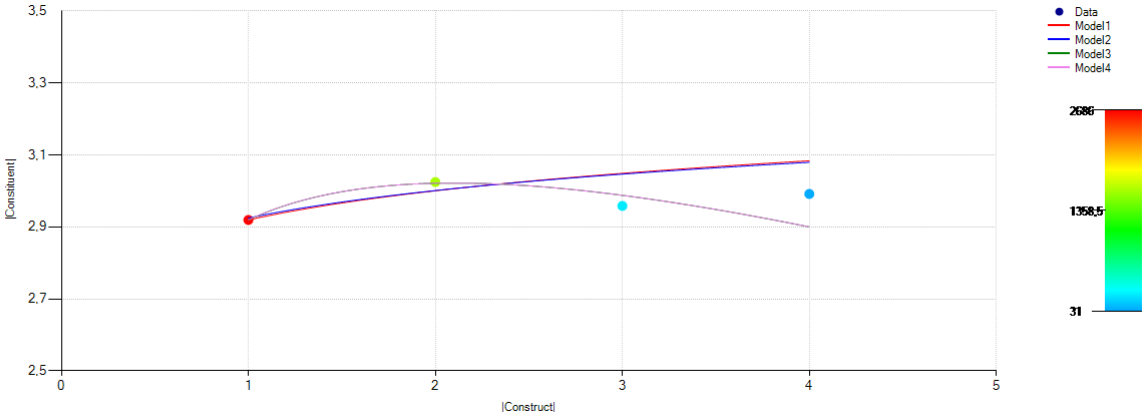
guómínshēngchǎnzǒngzhí（国民生产总值）

jìsuànjītǐcéngchéngxiàngyí（计算机体层成像仪）

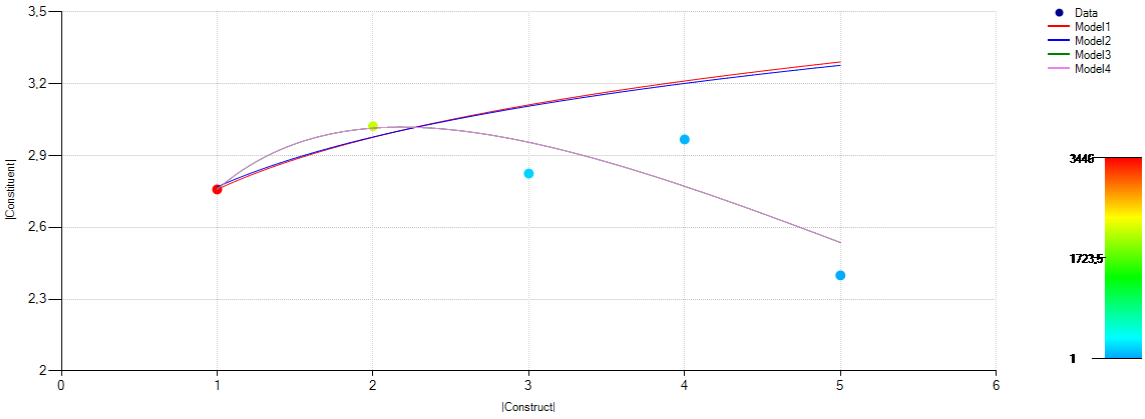
shìjièfēiwùzhìwénhuàyíchǎn（世界非物质文化遗产）

3. Grafické vizualizace, parametry *A* a *b* a koeficienty determinace R^2 pro všechny modely MALu

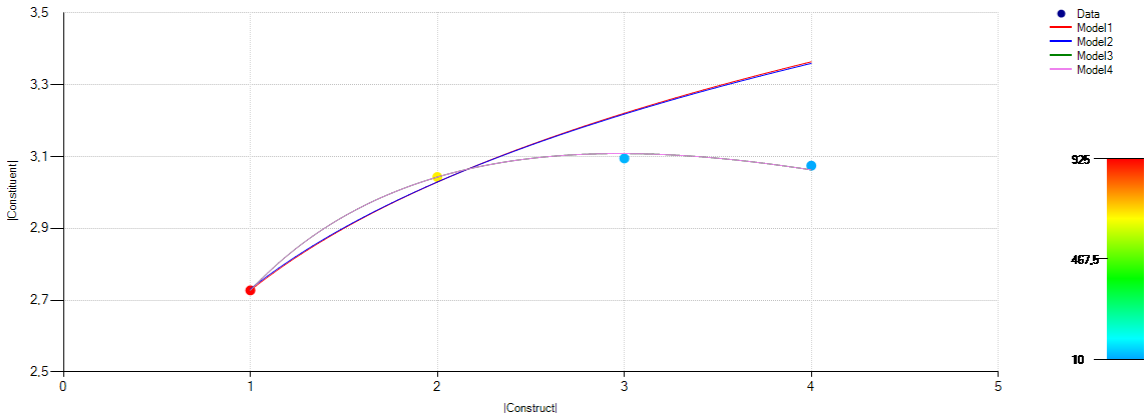
Jazyková úroveň U3



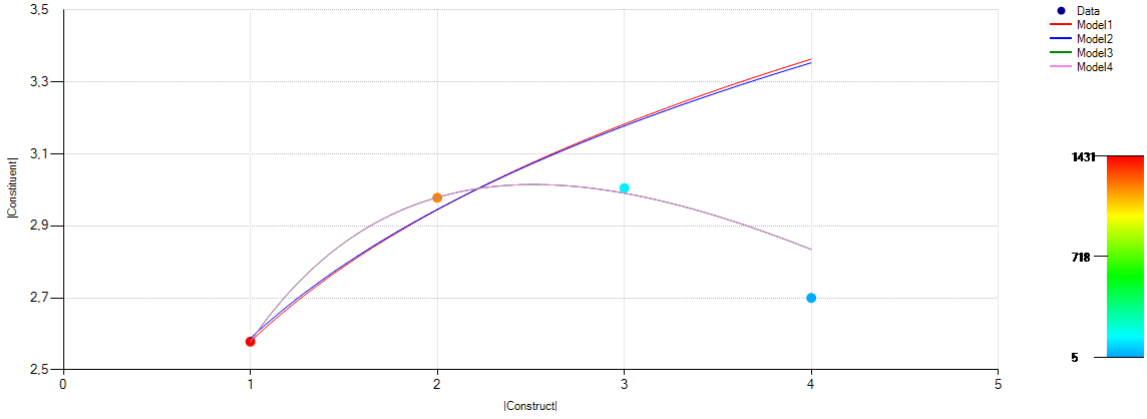
Obrázek 1 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 12 s proloženou regresní křivkou, Výběrový soubor 1: Yu Hua – *Kamarádi* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů)



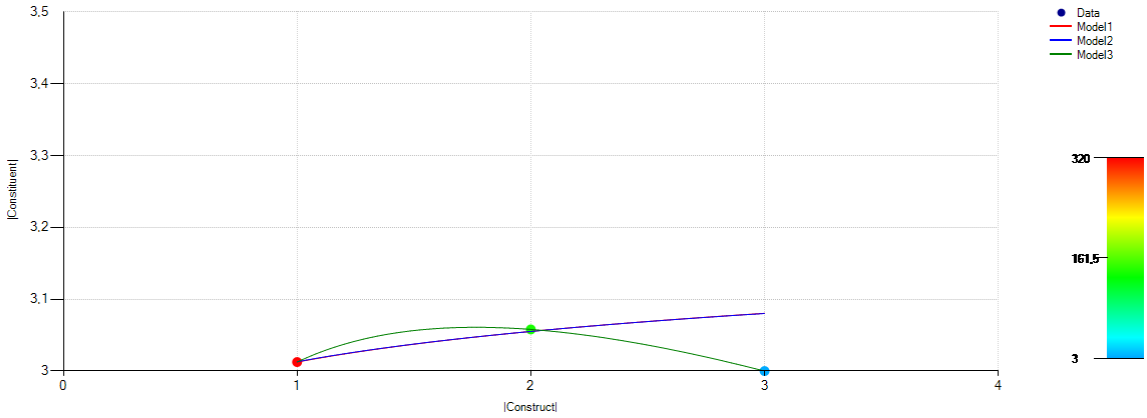
Obrázek 2 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 12 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů)



Obrázek 3 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 12 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů)



Obrázek 4 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 12 s proloženou regresní křivkou, Výběrový soubor 4: Zhang Liqing – *Deníkové záznamy v pinyinu* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů)

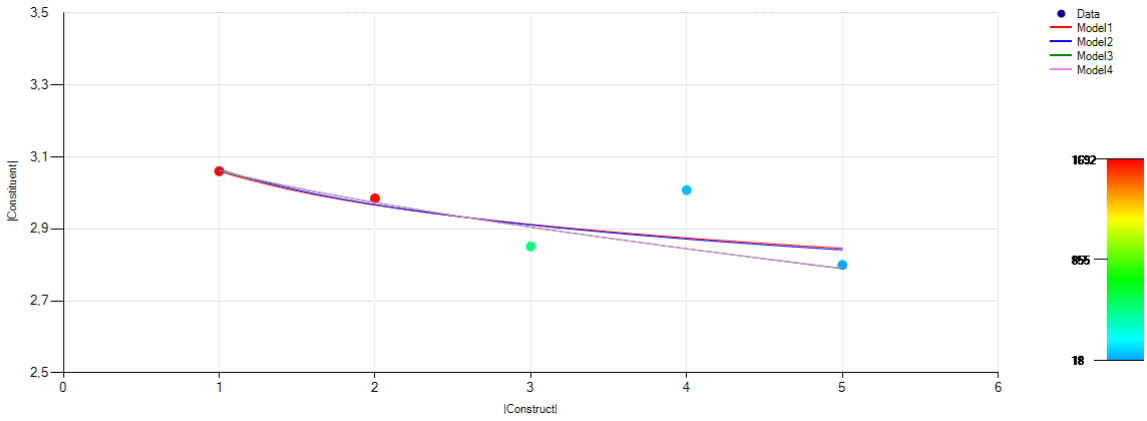


Obrázek 5 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 12 s proloženou regresní křivkou, Výběrový soubor 5: *Integrated Chinese Level 1 Part 2* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů)

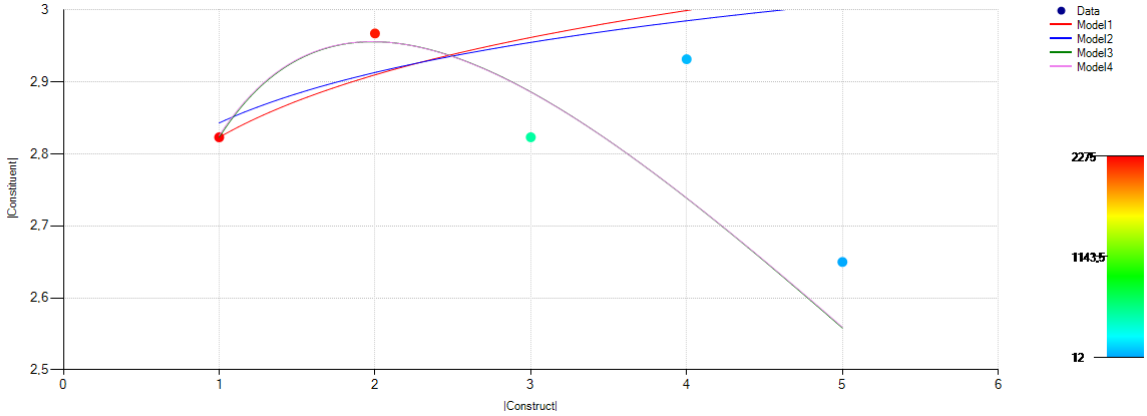
Tabulka 1 – Přílohy Parametry *A*, *b*, *c* a koeficient determinace R^2 pro všechny matematické modely MAL vztahující se k empiricky získaným datům na **úrovni U3**

Výběrový soubor	**Model MAL**	**Parametr *A***	**Parametr *b***	**Parametr *c***	***Koeficient determinace* R^2**
Výběrový soubor 1	Model 1	–	-0,04	–	0,84
	Model 2	2,93	-0,04	–	0,74
	Model 3	–	-0,16	-0,08	0,97
	Model 4	3,15	-0,16	-0,08	0,96
Výběrový soubor 2	Model 1	–	-0,11	–	0,87
	Model 2	2,77	-0,10	–	0,79
	Model 3	–	-0,38	-0,17	0,97
	Model 4	3,28	-0,38	-0,17	0,95
Výběrový soubor 3	Model 1	–	-0,15	–	0,98
	Model 2	2,73	-0,15	–	0,97
	Model 3	–	-0,31	-0,10	1,00
	Model 4	3,02	-0,31	-0,10	1,00
Výběrový soubor 4	Model 1	–	-0,19	–	0,97
	Model 2	2,59	-0,19	–	0,95
	Model 3	–	-0,49	-0,19	1,00
	Model 4	3,13	-0,49	-0,19	1,00
Výběrový soubor 5	Model 1	–	-0,02	–	0,93
	Model 2	3,01	-0,02	–	0,90
	Model 3	–	-0,12	-0,07	1,00
	Model 4	–	–	–	–

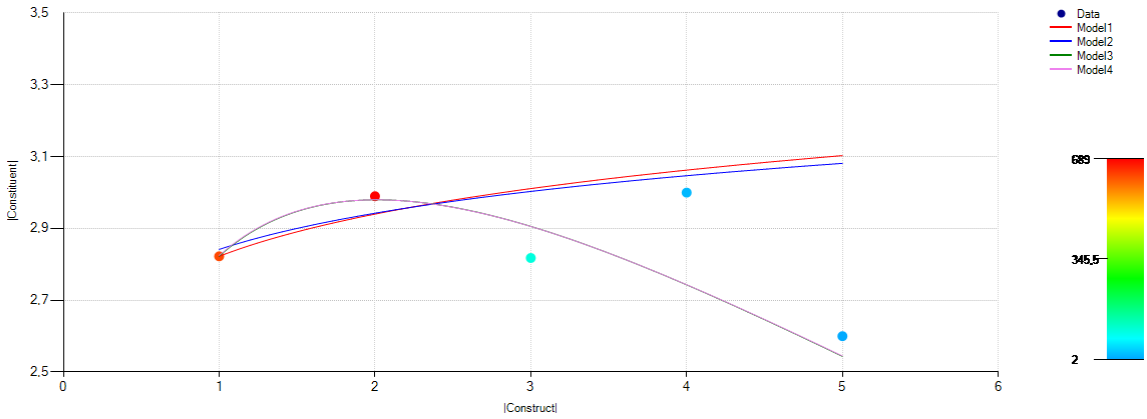
Jazyková úroveň U3 – Experiment 1



Obrázek 6 – Přílohy Experiment 1: Grafická vizualizace pozorování uvedených v Tabulce 16 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů) pro námi navržený způsob segmentace



Obrázek 7 – Přílohy Experiment 1: Grafická vizualizace pozorování uvedených v Tabulce 16 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů) pro námi navržený způsob segmentace

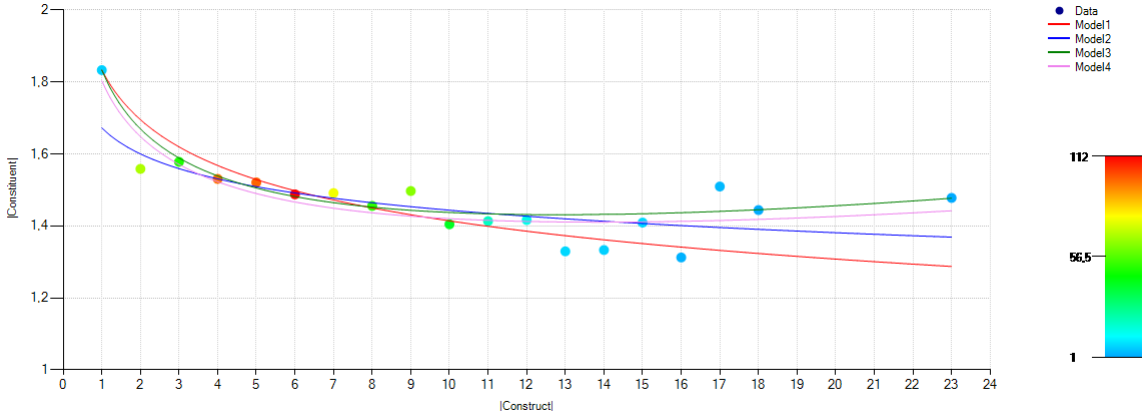


Obrázek 8 – Přílohy Experiment 1: Grafická vizualizace pozorování uvedených v Tabulce 16 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U3 (slovo měřené ve slabikách – slabika měřená v průměrném počtu grafémů) pro námi navržený způsob segmentace

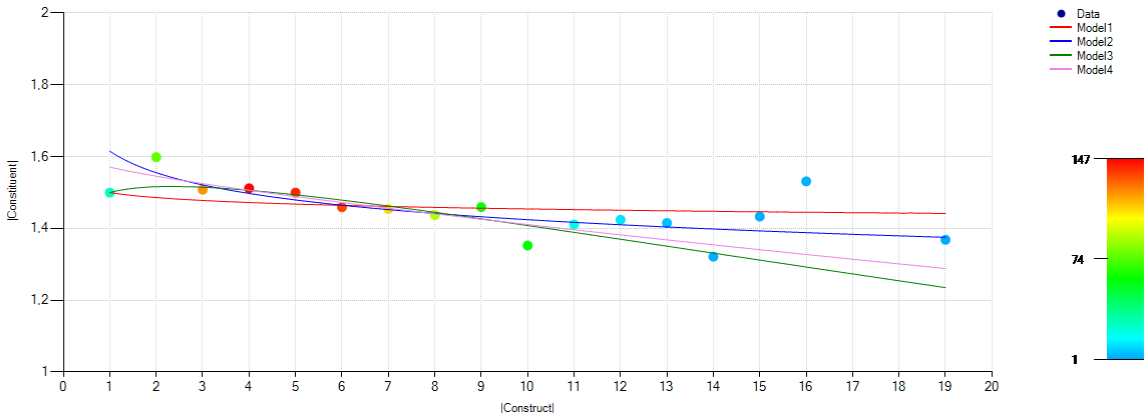
Tabulka 2 – Přílohy Experiment 1: Parametry *A*, *b*, *c* a koeficient determinace R^2 pro všechny matematické modely MAL vztahující se k empiricky získaným datům na **úrovni U3**

Výběrový soubor	**Model MAL**	**Parametr *A***	**Parametr *b***	**Parametr *c***	***Koeficient determinace R^2***
Výběrový soubor 1	Model 1	–	0,05	–	0,88
	Model 2	3,07	0,05	–	0,80
	Model 3	–	0,02	-0,02	0,89
	Model 4	3,11	0,02	-0,01	0,81
Výběrový soubor 2	Model 1	–	-0,04	–	0,61
	Model 2	2,84	-0,04	–	0,34
	Model 3	–	-0,24	-0,12	0,90
	Model 4	3,19	-0,24	-0,12	0,83
Výběrový soubor 3	Model 1	–	-0,06	–	0,71
	Model 2	2,84	-0,05	–	0,46
	Model 3	–	-0,28	-0,14	0,91
	Model 4	3,24	-0,27	-0,14	0,83

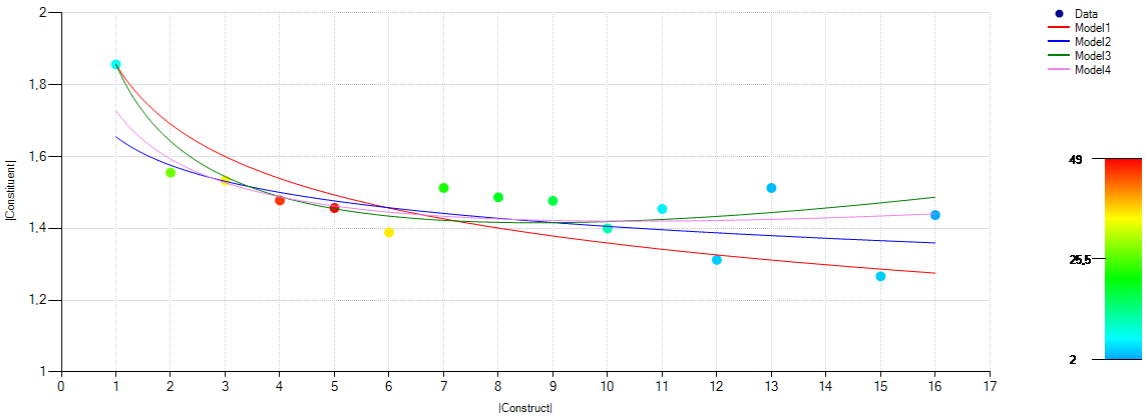
Jazyková úroveň U2



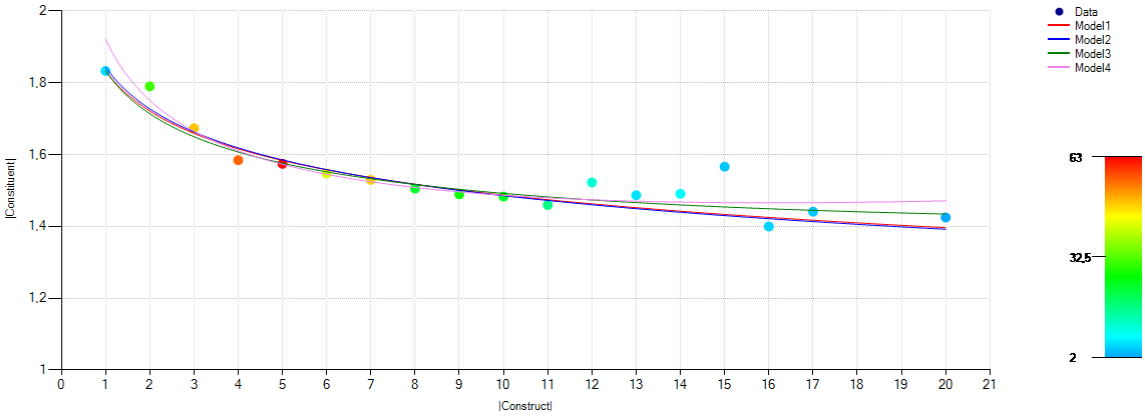
Obrázek 9 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 24 s proloženou regresní křivkou, Výběrový soubor 1: Yu Hua – *Kamarádi* pro jazykovou úroveň U2 (klauze měřená ve slovech – slovo měřené v průměrném počtu slabik)



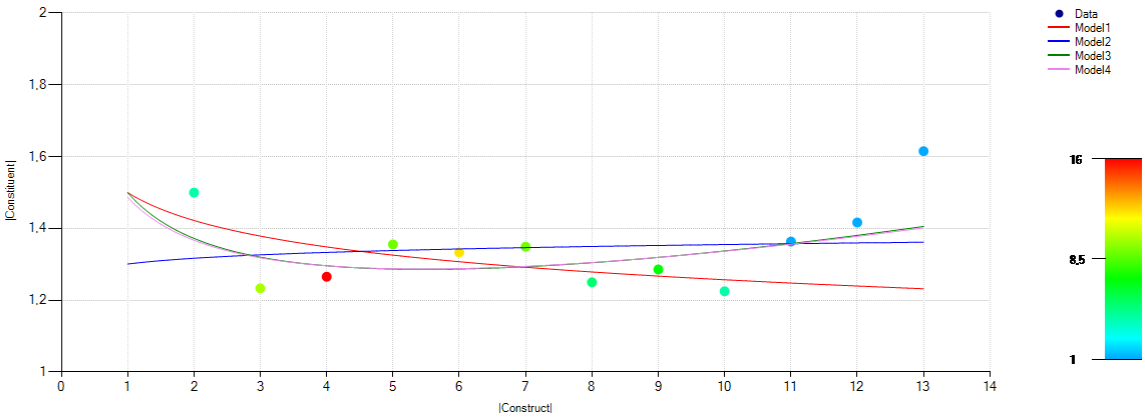
Obrázek 10 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 24 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U2 (klauze měřená ve slovech – slovo měřené v průměrném počtu slabik)



Obrázek 11 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 24 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U2 (klauze měřené ve slovech – slovo měřené v průměrném počtu slabik)



Obrázek 12 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 24 s proloženou regresní křivkou, Výběrový soubor 4: Zhang Liqing – *Deníkové záznamy v pinyinu* pro jazykovou úroveň U2 (klauze měřená ve slovech – slovo měřené v průměrném počtu slabik)

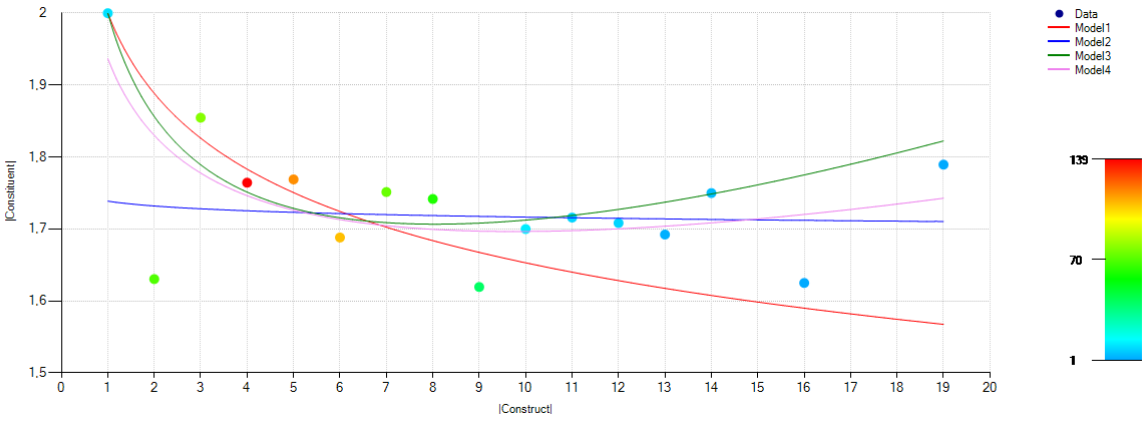


Obrázek 13 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 24 s proloženou regresní křivkou, Výběrový soubor 5: Integrated Chinese Level 1 Part 2 pro jazykovou úroveň U2 (klauze měřená ve slovech – slovo měřené v průměrném počtu slabik)

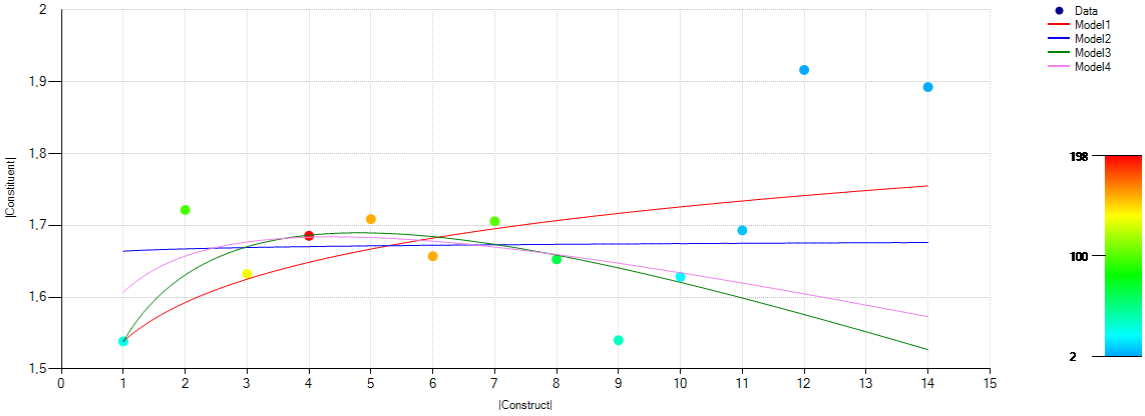
Tabulka 3 – Přílohy Parametry *A*, *b*, *c* a koeficient determinace R^2 pro všechny matematické modely MAL vztahující se k empiricky získaným datům na **úrovni U2**

Výběrový soubor	**Model MAL**	**Parametr *A***	**Parametr *b***	**Parametr *c***	***Koeficient determinace R^2***
Výběrový soubor 1	Model 1	–	0,11	–	0,97
	Model 2	1,67	0,06	–	0,74
	Model 3	–	0,15	0,01	0,98
	Model 4	1,79	0,14	0,01	0,88
Výběrový soubor 2	Model 1	–	0,01	–	0,34
	Model 2	1,61	0,06	–	0,65
	Model 3	–	-0,04	-0,02	0,68
	Model 4	1,59	0,01	-0,01	0,71
Výběrový soubor 3	Model 1	–	0,14	–	0,95
	Model 2	1,66	0,07	–	0,53
	Model 3	–	0,21	0,02	0,97
	Model 4	1,71	0,14	0,01	0,60
Výběrový soubor 4	Model 1	–	0,09	–	0,99
	Model 2	1,84	0,09	–	0,89
	Model 3	–	0,10	0,00	0,99
	Model 4	1,90	0,15	0,01	0,93
Výběrový soubor 5	Model 1	–	0,08	–	0,81
	Model 2	1,30	-0,02	–	0,02
	Model 3	–	0,17	0,03	0,88
	Model 4	1,44	0,16	0,03	0,12

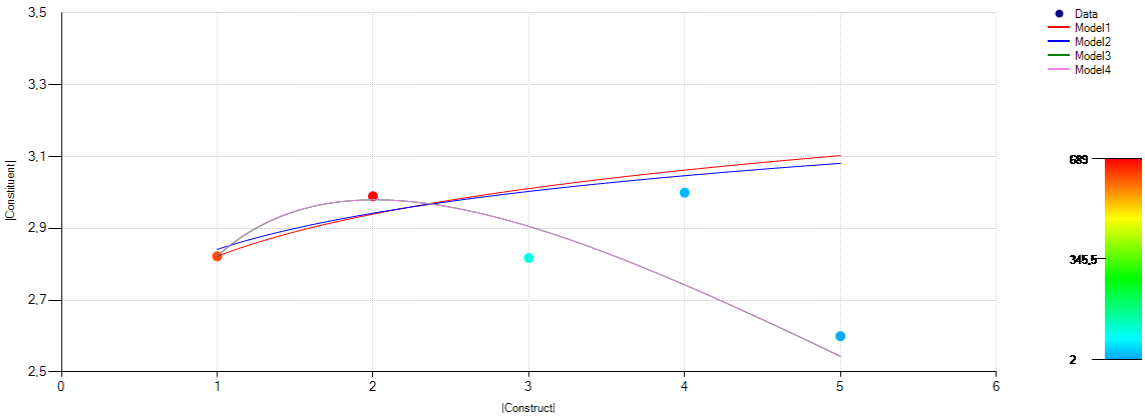
Jazyková úroveň U2 – Experiment 2



Obrázek 14 – Přílohy Experiment 2: Grafická vizualizace pozorování uvedených v Tabulce 27 s proloženou regresní křivkou, Výběrový soubor 1: Yu Hua – *Kamarádi* pro jazykovou úroveň U2 (klauze – slovo) pro námi navržený způsob segmentace



Obrázek 15 – Přílohy Experiment 2: Grafická vizualizace pozorování uvedených v Tabulce 27 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U2 (klauze – slovo) pro námi navržený způsob segmentace

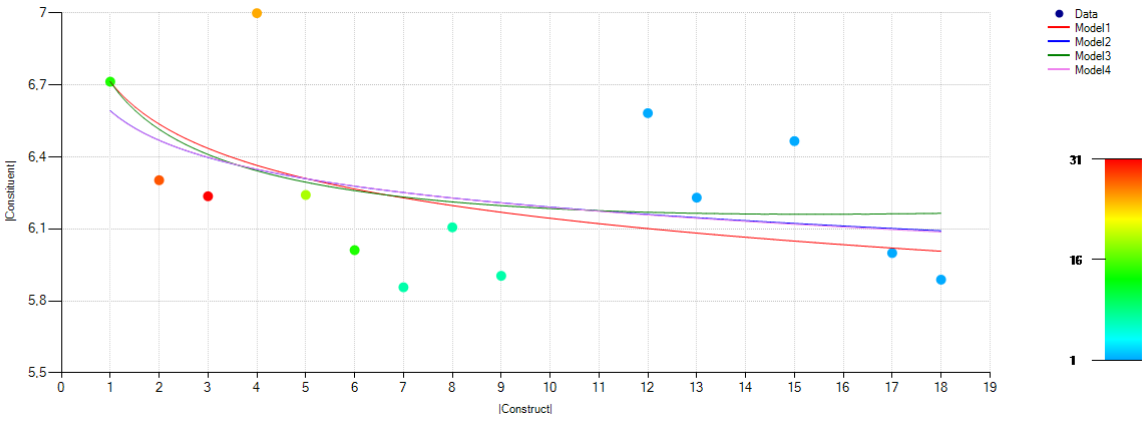


Obrázek 16 – Přílohy Experiment 2: Grafická vizualizace pozorování uvedených v Tabulce 27 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U2 (klauze – slovo) pro námi navržený způsob segmentace

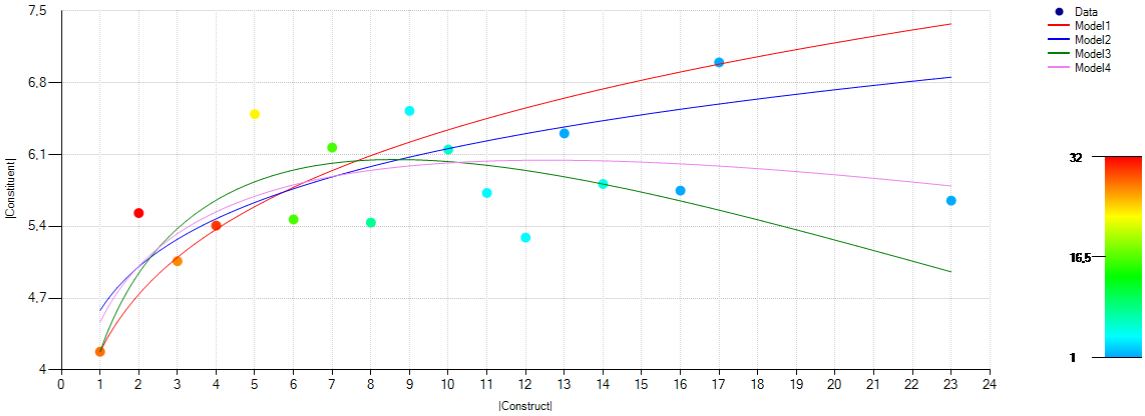
Tabulka 4 – Přílohy Experiment 2: Parametry *A*, *b*, *c* a koeficient determinace R^2 pro všechny matematické modely MAL vztahující se k empiricky získaným datům na **úrovni U2**

Výběrový soubor	**Model MAL**	**Parametr *A***	**Parametr *b***	**Parametr *c***	***Koeficient determinace R^2***
Výběrový soubor 1	Model 1	–	0,08	–	0,88
	Model 2	1,74	0,01	–	0,01
	Model 3	–	0,13	0,02	0,90
	Model 4	1,92	0,10	0,01	0,21
Výběrový soubor 2	Model 1	–	-0,05	–	0,82
	Model 2	1,66	0,00	–	1,14
	Model 3	–	-0,12	-0,03	1,13
	Model 4	1,63	-0,07	-0,02	1,06
Výběrový soubor 3	Model 1	–	0,12	–	0,88
	Model 2	1,75	0,03	–	0,11
	Model 3	–	0,25	0,05	0,95
	Model 4	1,77	0,14	0,03	0,29

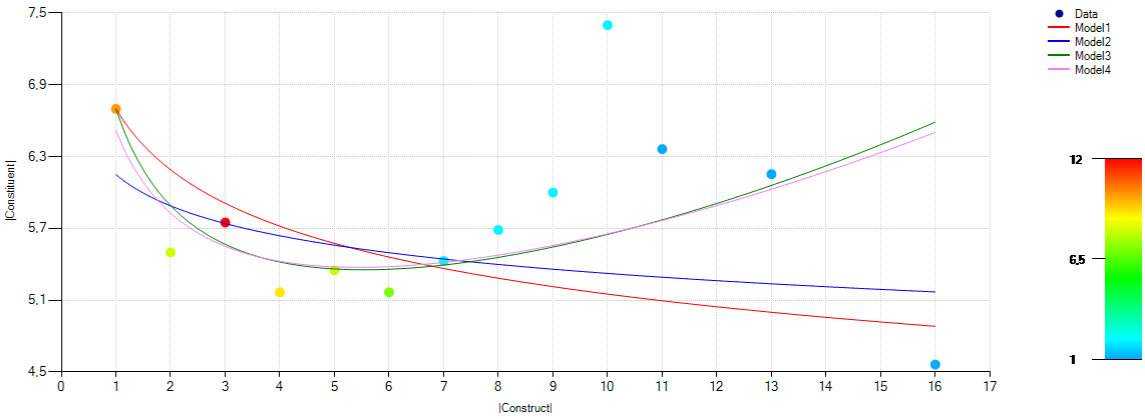
Jazyková úroveň U1



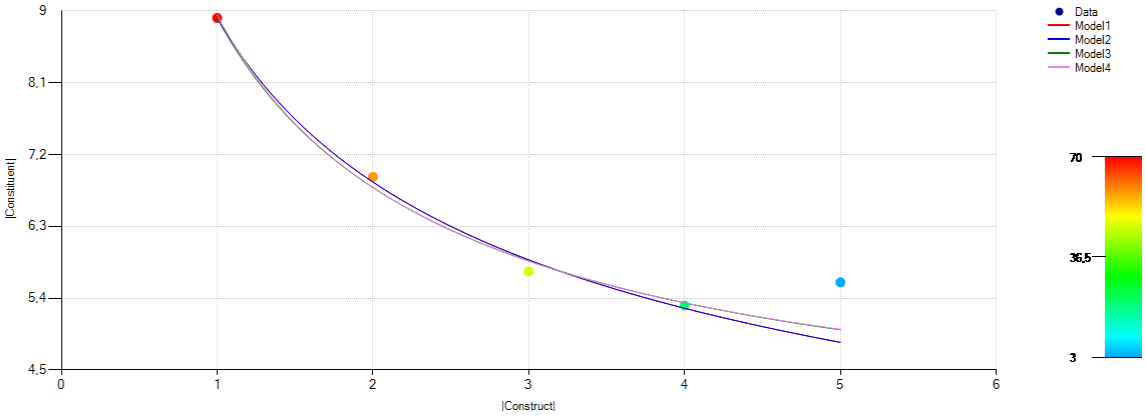
Obrázek 17 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 29 s proloženou regresní křivkou, Výběrový soubor 1: Yu Hua – *Kamarádi* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřená v průměrném počtu slov)



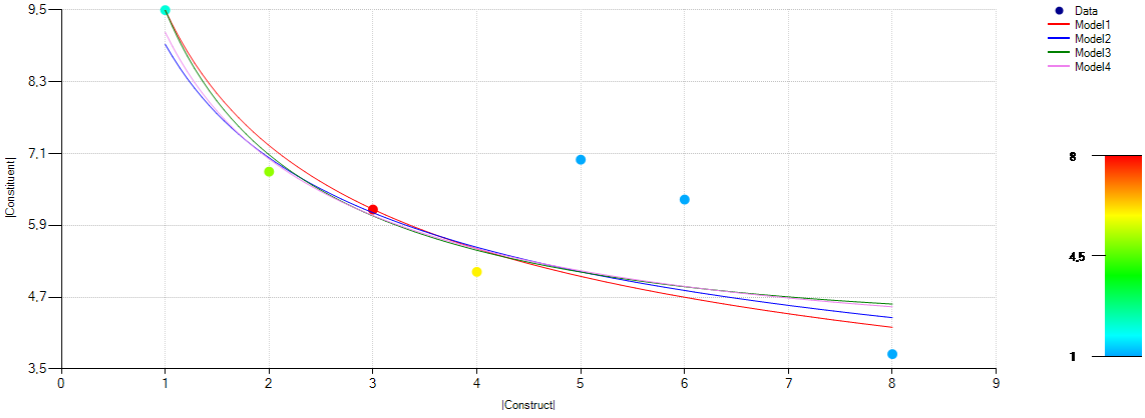
Obrázek 18 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 29 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřená v průměrném počtu slov)



Obrázek 19 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 29 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřená v průměrném počtu slov)



Obrázek 20 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 29 s proloženou regresní křivkou, Výběrový soubor 4: Zhang Liqing – *Deníkové záznamy v pinyinu* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřená v průměrném počtu slov)

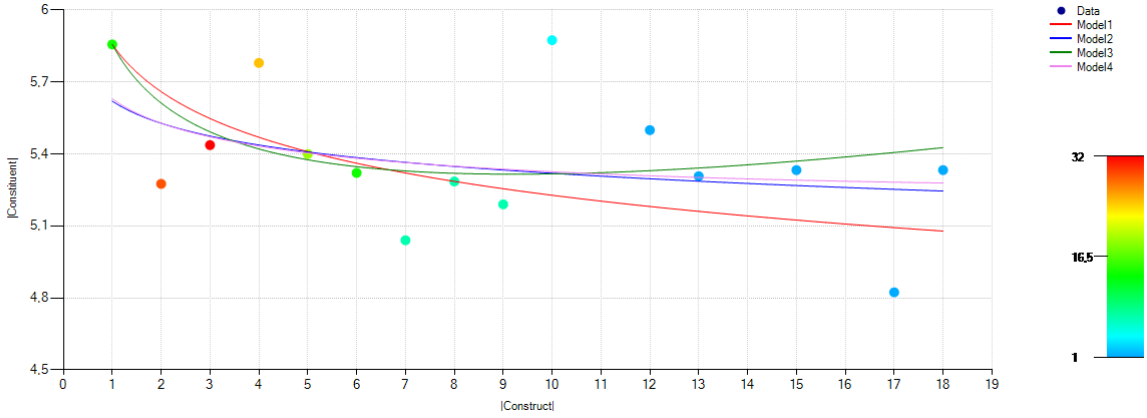


Obrázek 21 – Přílohy Grafická vizualizace pozorování uvedených v Tabulce 29 s proloženou regresní křivkou, Výběrový soubor 5: *Integrated Chinese Level 1 Part 2* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřené v průměrném počtu slov)

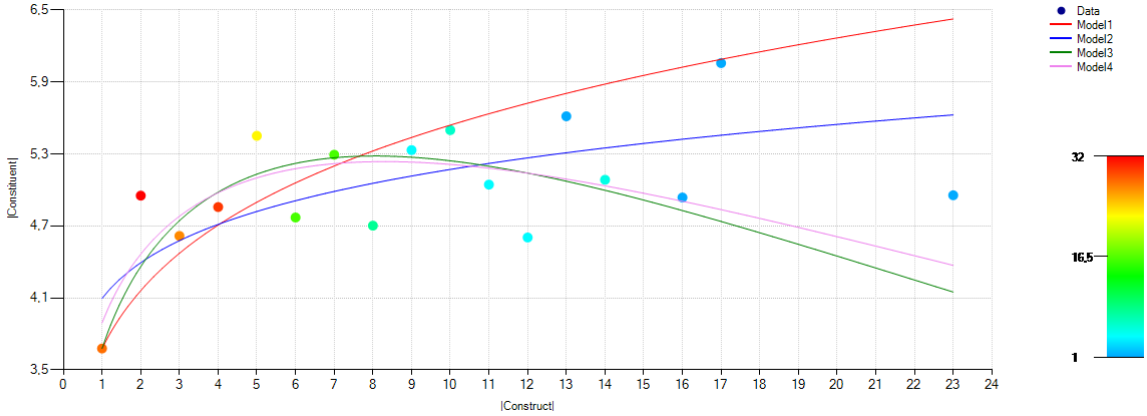
Tabulka 5 – Přílohy Parametry *A*, *b*, *c* a koeficient determinace R^2 pro všechny matematické modely MAL vztahující se k empiricky získaným datům na **úrovni U1**

Výběrový soubor	**Model**	**Parametr *A***	**Parametr *b***	**Parametr *c***	***Koeficient determinace R^2***
Výběrový soubor 1	Model 1	–	0,04	–	0,53
	Model 2	6,59	0,03	–	0,10
	Model 3	–	0,05	0,00	0,54
	Model 4	6,59	0,03	0,00	0,10
Výběrový soubor 2	Model 1	–	-0,18	–	0,89
	Model 2	4,58	-0,13	–	0,56
	Model 3	–	-0,29	-0,03	0,93
	Model 4	4,53	-0,19	-0,02	0,40
Výběrový soubor 3	Model 1	–	0,11	–	0,73
	Model 2	6,15	0,06	–	0,19
	Model 3	–	0,25	0,05	0,84
	Model 4	6,26	0,22	0,04	0,44
Výběrový soubor 4	Model 1	–	0,38	–	0,99
	Model 2	8,91	0,38	–	0,99
	Model 3	–	0,44	0,03	1,00
	Model 4	8,63	0,45	0,04	0,99
Výběrový soubor 5	Model 1	–	0,39	–	0,95
	Model 2	8,93	0,34	–	0,73
	Model 3	–	0,48	0,04	0,96
	Model 4	8,92	0,42	0,02	0,73

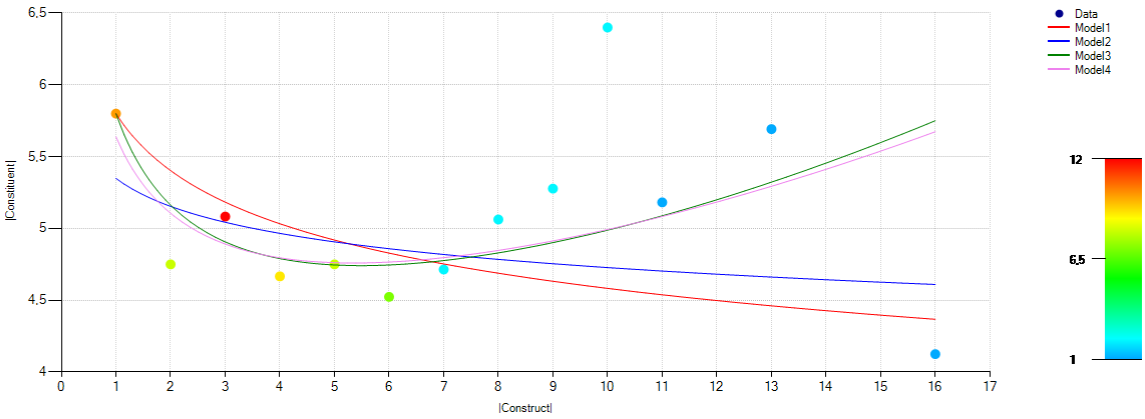
Jazyková úroveň U1 – Experiment 3



Obrázek 22 – Přílohy Experiment 3: Grafická vizualizace pozorování uvedených v Tabulce 31 s proloženou regresní křivkou, Výběrový soubor 1: Yu Hua – *Kamarádi* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřená v průměrném počtu slov) pro námi navržený způsob segmentace



Obrázek 23 – Přílohy Experiment 3: Grafická vizualizace pozorování uvedených v Tabulce 31 s proloženou regresní křivkou, Výběrový soubor 2: Yu Hua – *Vítězství ženy* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřená v průměrném počtu slov) pro námi navržený způsob segmentace



Obrázek 24 – Přílohy Experiment 3: Grafická vizualizace pozorování uvedených v Tabulce 31 s proloženou regresní křivkou, Výběrový soubor 3: Han Han – *Život, jak mu rozumím* pro jazykovou úroveň U1 (souvětí měřené v klauzích – klauze měřené v průměrném počtu slov) pro námi navržený způsob segmentace

Tabulka 6 – Přílohy Experiment 3: Parametry *A*, *b*, *c* a koeficient determinace R^2 pro všechny matematické modely MAL vztahující se k empiricky získaným datům na **úrovni U1**

Výběrový soubor	**Model MAL**	**Parametr *A***	**Parametr *b***	**Parametr *c***	***Koeficient determinace R^2***
Výběrový soubor 1	Model 1	–	0,05	–	0,74
	Model 2	5,62	0,02	–	0,68
	Model 3	–	0,07	0,01	0,71
	Model 4	5,63	0,03	0,00	0,67
Výběrový soubor 2	Model 1	–	-0,18	–	0,89
	Model 2	4,10	-0,10	–	0,54
	Model 3	–	-0,30	-0,04	0,94
	Model 4	4,01	-0,24	-0,03	0,67
Výběrový soubor 3	Model 1	–	0,10	–	0,71
	Model 2	5,35	0,05	–	0,17
	Model 3	–	0,23	0,04	0,83
	Model 4	5,44	0,20	0,04	0,37