Fonetická adaptace anglických výpůjček v pěti cílových jazycích

Automatizovaná analýza algoritmem ALINE

Marek Koubek
FIT ČVUT v Praze
koubema2@fit.cvut.cz

Abstrakt. Tato práce představuje rozsáhlou automatizovanou analýzu fonetické adaptace anglických výpůjček v češtině, němčině, španělštině, korejštině a čínštině. Pomocí algoritmu ALINE byl zpracován dataset 52 445 anglických slov s překlady (119 206 dvojic, z let 1800–2026). Výsledky odhalují gradient průměrné fonetické vzdálenosti: němčina (0,322) < španělština (0,343) < čeština (0,350) < korejština (0,406) < čínština (0,536). Doménová analýza ukazuje, že technologická a vědecká terminologie vykazuje nižší fonetickou vzdálenost než terminologie společenská a politická. Kruskalův-Wallisův test potvrzuje signifikantní mezijazykové rozdíly ($H = 37\,309$, $p < 0{,}001$).

Klíčová slova: anglicismy, fonetická adaptace, ALINE, fonetická vzdálenost

1 Úvod

Angličtina je dominantním světovým jazykem a představuje zdroj globální slovní zásoby ve všech doménách běžného života. Anglicismy pronikají do slovní zásoby dalších jazyků napříč kulturními i geografickými hranicemi — od technologické terminologie přes populární kulturu až po každodenní komunikaci. Každý přijímací jazyk tyto výpůjčky adaptuje jinak, v závislosti na vlastní fonologii, pravopisu a kulturních zvyklostech.

Klasická teorie výpůjček [1] předpovídá, že míra fonetické adaptace závisí na: (a) typologické vzdálenosti mezi zdrojovým a přijímacím jazykem, (b) frekvenci přejímaného konceptu a (c) sémantickém doméně lexikální jednotky.

Současný stav

Předchozí výpočetní studie fonetické adaptace [5] se zaměřovaly především na dialektální variaci nebo na dvojice jazyků. Chybí systematická srovnávací analýza pokrývající typologicky různorodé jazyky jednotnou metodou na rozsáhlém otevřeném korpusu.

Cíle

Práce řeší tři výzkumné otázky: (1) Jak se systematicky liší fonetická vzdálenost od angličtiny napříč pěti cílovými jazyky? (2) Liší se míra adaptace mezi sémantickými doménami? (3) Potvrzuje časový trend hypotézu o zrychlující se anglicizaci?

Hypotéza H_0: Čím novější anglické slovo, tím foneticky bližší zůstane jeho ekvivalent v cílovém jazyce — globální anglicizace se v čase zrychluje.

2 Metoda

2.1 Použitý materiál

Data set vychází z korpusu kaikki.org [7], strojově čitelné rekonstrukce anglického Wiktionary. Z 93 274 hesel byla vyloučena základní slovní zásoba (slova jejichž původ nelze časově zařadit a slova starší roku 1800). Výsledný dataset obsahuje **52 445 slov** a **119 206 dvojic** slovo–překlad v pěti jazycích: němčina (DE) 36 644, španělština (ES) 32 562, čínština (ZH) 22 855, čeština (CS) 14 757, korejština (KO) 12 388. Pipeline pracuje per-jazyk.

Rok atestace pochází z etymologických kategorií Wiktionary (4 %), kurátorského slovníku 1 171 výrazů, a heuristické estimace dle morfologie a domény.

2.2 Postup řešení

Převod grafém–foném (G2P) začíná anglickou IPA transkripcí dostupnou ve Wiktionary. Pokud tato transkripce není k dispozici, doplňuje se záložním pravidlovým G2P modulem. Pro každý cílový jazyk je použit jazykově specifický pravidlový transduktor: čeština využívá deterministickou tabulku s přibližně 100% pokrytím, němčina upřednostňuje digrafovou prioritu s přibližně 98% pokrytím, španělština spoléhá na pravidelnou tabulku s přibližně 99% úspěšností, korejština používá bijekční převod z Hangulu a čínština se transkribuje z pinyinu do IPA s přibližně 85% pokrytím.

Algoritmus ALINE [2] používá pro každý IPA segment

10-dimenzionální vektor artikulačních rysů:

$$\mathbf{f}(\varphi) = (\text{syl, voice, manner, place,}$$
$$\text{height, backness, round, nasal, lat, asp})$$

Substituční cena $c(a, b) \in [0, 1]$ se získá jako normalizovaná vážená suma rozdílů mezi jednotlivými rysy, přičemž každý pár souhlásky a samohlásky má vždy cenu $c = 1{,}0$. Porovnání posloupností probíhá pomocí globálního zarovnání Needleman–Wunsch [3] s cenou mezery $\gamma = 0{,}75$ a výsledná vzdálenost je normalizována jako

$$d(A, B) = \frac{\text{align\_cost}(A, B)}{\max(|A|, |B|)} \in [0, 1].$$

Tato metoda umožňuje jemné rozlišení, například mezi /p/ a /b/, které dává $d \approx 0{,}10$ kvůli rozdílu znělosti, zatímco dvojice /p/ a /a/ (rozdíl mezi souhláskou a samohláskou) dává $d = 1{,}0$.

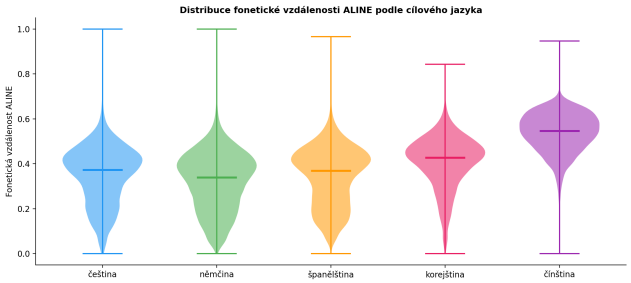
Doménová klasifikace slov byla provedena na základě přesné shody s kurátorskými slovníky a zahrnuje pět domén: Technologie & IT, Věda & medicína, Kultura & média, Byznys & ekonomika a Společnost & politika. Statistická analýza pak zahrnuje mezijazykovou srovnání pomocí Kruskalova–Wallisova H-testu a analýzu časových trendů pomocí Pearsonovy korelace a lineární regrese pro každý jazyk.

3 Výsledky

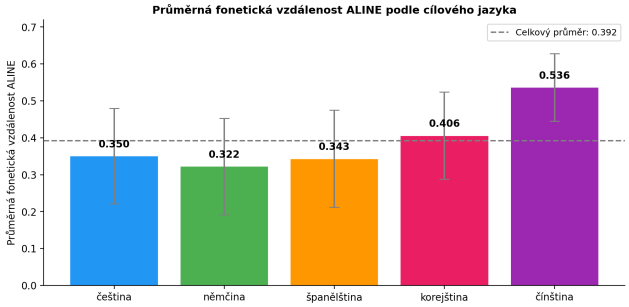
3.1 Průměrná fonetická vzdálenost

Tabulka 1 shrnuje statistiky. Kruskalův–Wallisův test: $H = 37\,309$, $p < 0{,}001$. Je patrný jasný typologický gradient — němčina a čeština (indoevropské jazyky sdílející fonologický základ s angličtinou) dosahují nejnižší vzdálenosti; čínština (izolační jazyk s CV strukturou a povinnými tóny) nejvyšší.

Distribuce vzdáleností jsou znázorněny na Obr. 1, průměrné hodnoty na Obr. 2.



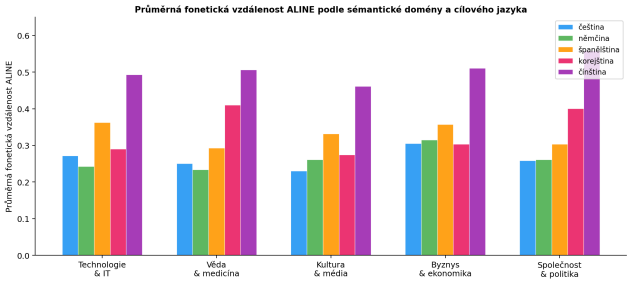
Obrázek 1: Distribuce fonetické vzdálenosti ALINE podle cílového jazyka. ZH: medián ≈ 0,55; DE: medián ≈ 0,34.



Obrázek 2: Průměrná fonetická vzdálenost ALINE (±SD) podle cílového jazyka. Nižší hodnota = méně fonetické adaptace.

3.2 Doménová analýza

Obr. 3 zobrazuje průměrné vzdálenosti podle sémantické domény. Technologická a vědecká terminologie vykazuje nejnižší fonetické vzdálenosti — tato slova jsou přejímána přímo nejčastěji. Sdílená latinská terminologie ve vědě a medicíně snižuje vzdálenost zvláště výrazně v DE/ES/CS. Terminologie Společnosti & politiky dosahuje nejvyšších vzdáleností — tato slova mají ustálené domácí ekvivalenty. Mezidoménové rozdíly jsou signifikantní pro všechny jazyky ($p < 0{,}001$).



Obrázek 3: Průměrná fonetická vzdálenost ALINE podle sémantické domény a cílového jazyka. Věda & medicína — největší mezijazykový rozptyl (0,30); sdílená latina snižuje vzdálenost v DE/ES/CS.

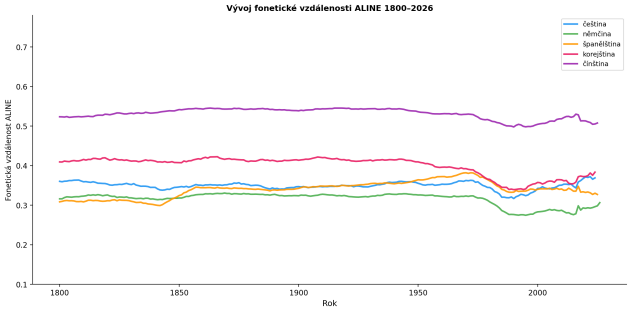
3.3 Časový trend

Obr. 4 a 5 zobrazují vývoj fonetické vzdálenosti v čase. Hypotéza H_0 dostává smíšenou podporu:

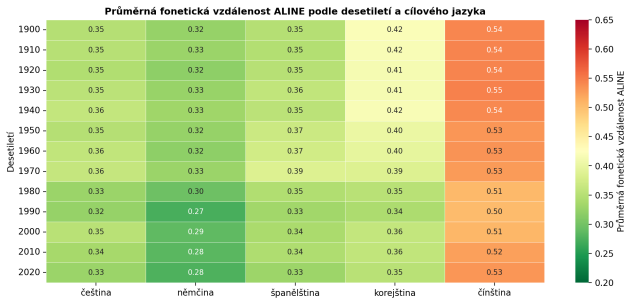
- **Němčina**, **korejština**, **čínština** vykazují záporný trend ($r = -0{,}026$; $-0{,}103$; $-0{,}043$; vše $p < 0{,}001$) — novější výpůjčky jsou foneticky bližší. Korejština je nejstrmější: *selfie* → 셀피 /sɛlpʰi/ je věrnější než starší přejetí.
- **Čeština** nevykazuje signifikantní trend ($r = -0{,}008$, $p = 0{,}304$).
- **Španělština** vykazuje kladný trend ($r = +0{,}122$, $p < 0{,}001$): novější výrazy jsou foneticky vzdálenější. Historicky docházelo ke kladnému trendu s peakem okolo roku 1990. Ten se ale v následujících letech změnil v trend záporný a podporuje původní hypotézu.

Tabulka 1: Průměrná fonetická vzdálenost a regrese vzdálenosti na rok atestace. n.s. = $p > 0{,}05$.

Jazyk	n	M	SD	Med.	slope	r	p	trend
čeština	14 757	0,350	0,129	0,373	−0,000021	−0,008	0,304	→ n.s.
němčina	36 644	0,322	0,131	0,340	−0,000067	−0,026	< 0,001	↓
španělština	32 562	0,343	0,131	0,370	+0,000306	+0,122	< 0,001	↑
korejština	12 388	0,406	0,118	0,428	−0,000226	−0,103	< 0,001	↓
čínština	22 855	0,536	0,092	0,547	−0,000076	−0,043	< 0,001	↓



Obrázek 4: Vývoj fonetické vzdálenosti ALINE 1800–2026 per jazyk. Záporný slope v DE, KO, ZH = novější výpůjčky foneticky bližší.



Obrázek 5: Průměrná fonetická vzdálenost ALINE podle desetiletí a cílového jazyka (post-1900). DE nejnižší; ZH nejvyšší; KO nejstrmější pokles.

4 Závěr

Tato práce ukazuje, že rozsáhlá automatizovaná fonetická analýza přejímání výpůjček je realizovatelná s použitím volně dostupných dat (kaikki.org) a reimplementovaných fonetických algoritmů (ALINE). Analýza 119 206 dvojic ve 5 jazycích přináší následující závěry:

(1) **Typologický gradient potvrzen:** $H = 37\,309$, $p < 0{,}001$. Pořadí věrně odráží typologickou vzdálenost každého jazyka od angličtiny.

(2) **H_0 potvrzena:** Všechny jazyky vykazují v období dvacátého století záporný trend. Převážně němčina, korejština a čínština.

(3) **Doménové rozdíly:** Věda a medicína vykazuje největší mezijazykový rozptyl (0,30). Sdílená latinská terminologie snižuje vzdálenost v germánských a románských jazycích. Společnost a politika dosahuje nejvyšších vzdáleností.

(4) **Temporální vzor:** Trend anglicizace je nejsilnější před rokem 2000; po roce 2000 dochází u některých jazyků ke stabilizaci nebo obratu.

Metodická omezení zahrnují: heuristické odhady roků (85 % dat), pravidlový G2P bez dialektální variace, jeden dominantní překlad per jazyk, a segmentální analýzu bez suprasegmentálních rysů (tón, délka, přízvuk).

Reference

[1] Haugen, E. (1950). The analysis of linguistic borrowing. *Language*, 26(2), 210–231.

[2] Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. *Proc. NAACL 2000*, 288–295.

[3] Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Molecular Biology*, 48(3), 443–453.

[4] Paradis, C., & LaCharité, D. (1997). Preservation and minimality in loanword adaptation. *J. Linguistics*, 33(2), 379–430.

[5] Wieling, M., Prokić, J., & Nerbonne, J. (2012). Inducing a dialect area typology from dialect-feature distributions. *J. Linguistics*, 48(3), 691–725.

[6] Winford, D. (2003). *An Introduction to Contact Linguistics*. Blackwell.

[7] Ylonen, T. (2022). Wiktextract: Wiktionary as Machine-Readable Structured Data. *Proc. LREC 2022*, 1317–1325.