# TEXTY V DIDAKTICKÝCH TESTECH Z PERSPEKTIVY KVANTITATIVNÍ SYNTAXE

## Didactic Test Texts from the Perspective of Quantitative Syntax

Michal Místecký – Lucie Radková – Žaneta Stiborská – Michaela Nogolová – Darina Hrubá

***Abstrakt:*** *Studie se zaměřuje na kvantitativněsyntaktickou analýzu textů, které slouží k testování čtenářské gramotnosti žáků v českých didaktických testech. Rozbory jsou prováděny na korpusu textů, jež byly excerpovány z didaktických testů určených k přijímacím zkouškám do osmiletých, šestiletých a čtyřletých středoškolských oborů a z didaktického testu, který je součástí maturitní zkoušky z českého jazyka, a to za období 2019–2023. Bylo zjišťováno, zdali jsou mezi testy statisticky významné rozdíly na základě frekvence široce pojatých atributů (atr_plus), příslovečných určení (adv), apozic (apos), jmenných částí přísudku (pnom), míry subordinace (subord), průměrné délky klauze počítané v lineárních dependenčních segmentech (ACL) a průměrného počtu slov tvořících jeden lineární dependenční segment (ALDSL). Statisticky signifikantní rozdíly se projevily u indikátorů atr_plus, adv, apos a ACL, přičemž základní dělicí osa leží mezi maturitními a přijímacími testy. Nakonec byla provedena analýza hlavních komponent, která poukázala na absenci přesně vymezitelných klastrů textů. První hlavní komponenta, jež vysvětluje 37 % celkové variability, nabývá vysokých hodnot pro texty syntakticky komplexní (s vysokými hodnotami atr_plus, apos, ACL a ALDSL) a nízkých pro ty, které preferují slovesnější struktury (s vysokými hodnotami adv a subord).*

***Klíčová slova:*** *syntax; didaktický test; kvantitativní lingvistika; analýza hlavních komponent*

***Abstract:*** *The study focuses on the quantitative-syntactic analysis of texts used to check students' reading literacy skills in Czech didactic tests. The analyses are carried out on a corpus of texts collected from didactic tests intended for the entrance exams to eight-year, six-year and four-year secondary studies, and from a didactic test that is part of the secondary school-leaving exam*

*in the Czech language ("maturita"). The covered period is 2019–2023. It is investigated whether there are statistically significant differences between the tests based on the frequency of broadly conceived attributes (atr_plus), adverbials (adv), appositions (apos), nominal parts of the predicate (pnom), degrees of subordination (subord), average lengths of clauses counted in linear dependency segments (ACL), and average numbers of words forming one linear dependency segment (ALDSL). Statistically significant differences are observed for the indicators of atr_plus, adv, apos and ACL, with the basic dividing line lying between the maturita and entrance exam tests. Finally, a principal component analysis (PCA) was conducted, which indicated the absence of clearly delineable text clusters. The first principal component, accounting for 37% of the total variance, takes on high values for syntactically complex texts (with high values of atr_plus, apos, ACL, and ALDSL) and low values for those favoring more verbal structures (with high values of adv and subord).*

***Keywords:*** *syntax; didactic test; quantitative linguistics; principal component analysis*

Korespondenční autoři: Michal Místecký, Filozofická fakulta, Ostravská univerzita, Reální 5, 701 03 Ostrava, e-mail: *michal.mistecky@osu.cz*; Lucie Radková, Filozofická fakulta, Ostravská univerzita, Reální 5, 701 03 Ostrava, e-mail: *lucie.radkova@osu.cz*

## 1. Úvod

V současné době dochází v lingvistice a didaktice českého jazyka ke dvěma paralelním trendům – jednak se rozněcuje veřejná i odborná debata nad kompozicí, zaměřením a smyslem didaktických testů z českého jazyka (srov. Štěpáník, 2018; Baumgartnerová, 2021–2022; též Kabátová, 2023; Münich, 2024; Čaban, 2024), které slouží jako přijímací zkoušky z této oblasti do příslušného cyklu středoškolského studia a v jiné formě jako testy maturitní, jednak se úspěšně rozvíjí potenciál syntaktických indikátorů jako markerů vypočitatelné složitosti textu (srov. Kubát a kol., 2021; Čech a kol., 2022; Kubát ‑ Čech, 2024; Kubát a kol., 2024). Oba tyto trendy lze zužitkovat pro obohacení probíhající diskuse o čtenářské gramotnosti, která je jedním z pilířů komunikačního pojetí výuky českého jazyka a literatury na současných českých školách (Štěpáník, 2020). Naše studie se proto bude zaměřovat na zhodnocení syntaktické komplexity vybraných textů, které jsou součástí didaktických testů; využije metodologie kvantitativní lingvistiky, jejímiž největšími

výhodami jsou snadná replikovatelnost a intersubjektivita výsledků, které poskytuje (srov. Davidson, 2004). Ze zřetele však nebudou vypouštěny ani didaktické souvislosti našich zjištění. Studie v tomto ohledu navazuje na obdobný výzkum, realizovaný ovšem z pohledu morfologie (Místecký a kol., 2025).

Na čtenářskou náročnost budeme nahlížet prizmatem současných lingvistických metod, konkrétně prostřednictvím zmíněného přístupu založeného výhradně na kvantitativních ukazatelích, a nikoli např. z perspektivy metakognitivních procesů při čtení (viz Liptáková – Cibáková, 2013). Propojení kvantitativní lingvistiky a pedagogického výzkumu představuje v českém kontextu dosud marginální, byť potenciálně slibnou oblast. V poslední době se začaly objevovat dílčí statisticky zakotvené studie věnované evaluaci školních slohových prací (srov. Místecký – Radková, 2020; Místecký – Radková, 2021–2022) či didaktice češtiny jako cizího jazyka (srov. Nogolová a kol., 2023a; Nogolová a kol., 2023b). Opomenout nelze ani dlouhou tradici matematizace jazykové složitosti textu, jejímiž průkopníky byli v tuzemsku Průcha (1998) a Pluskal (1996), kteří své metody ovšem většinově aplikovali na učebnice. Tyto metody se však podrobně nezaměřují na syntaktickou charakteristiku textů, která je předmětem našeho zájmu v tomto článku. V kontextu světové pedagogiky je ale výše zmíněné spojení zevrubně badatelsky vytěžováno, a jsou dokonce vyvíjeny i nástroje, které mají za cíl komplexitu textu na základě vybraných ukazatelů měřit (srov. Graesser a kol., 2011; Rafatbakhsh ‐ Ahmadi, 2023).

Studie je členěna do čtyř tematických celků. Po první, úvodní pasáži následuje část druhá, ve které představíme korpus textů, které podrobujeme analýze, počítané syntaktické indikátory a statistický aparát, jejž využijeme ke srovnání testových sad mezi sebou. Celkem budou provedeny dva rozbory – základní statistický a analýza hlavních komponent. Třetí část obsahuje výsledky výpočtů a rozborů a jejich interpretace. Ve čtvrté části poukazujeme na meze využitého metodologického aparátu a naznačujeme, jakým dalším směrem by se mohlo bádání v naší oblasti ubírat.

## 2. Materiál a metody

Za texty, jež využíváme v naší analýze, považujeme jazykové komunikáty, které jsou součástí zadání těch úloh v didaktických testech, jež se zjevně zaměřují na čtenářskou gramotnost. Vylučujeme texty krátké (N < 100) a syntakticky nesouvislé (např. s převládajícím podílem nevětných výčtů). Za jeden text považujeme odstavce, které je třeba sestavit do smysluplného celku, a charakteristiky autorů či literárních škol. Do výzkumu zahrnujeme didaktické testy, jež tvoří přijímací zkoušky z českého jazyka do osmiletých, šestiletých a čtyřletých středoškolských studijních oborů (tyto značíme „E8“, „E6“ a „E4“; E = „entrance exam“), a matu-

ritní didaktické testy z českého jazyka (značíme „M“). Testy pokrývají období let 2019–2023 a jsou volně dostupné na stránkách Centra pro zjišťování výsledků vzdělávání (Cermat, 2019); složení korpusu zachycuje tabulka 1. Každému textu je dále přidělen jedinečný kód, jehož strukturu přibližuje tabulka 2.

Tabulka 1
*Stavba zkoumaného korpusu v počtech textů a testů (v závorce).*

| | E8 | E6 | E4 | M | celkem |
|---|---|---|---|---|---|
| 2019 | 14 (2) | 13 (2) | 12 (2) | 12 (2) | 51 (8) |
| 2020 | 6 (1) | 6 (1) | 6 (1) | 14 (2) | 32 (5) |
| 2021 | 21 (4) | 23 (4) | 24 (4) | 20 (3) | 88 (15) |
| 2022 | 21 (4) | 23 (4) | 22 (4) | 13 (2) | 79 (14) |
| 2023 | 24 (4) | 23 (4) | 23 (4) | 14 (2) | 84 (14) |
| celkem | 86 (15) | 88 (15) | 87 (15) | 73 (11) | 334 (56) |

Tabulka 2
*Struktura kódu textu. U přijímacích zkoušek jsou termíny označovány jako řádný (1) a mimořádný (2) a u maturit jako mimořádný jarní (0), jarní (1) a podzimní (2).*

| test | typ testu | rok | termín | pořadí termínu | pořadí textu v testu | příklad |
|---|---|---|---|---|---|---|
| přijímací | E8/E6/E4 | 2019–2023 | 1/2 | 1/2 | 1–7 | E8_2019_1_1_2 |
| maturitní | M | 2019–2023 | 0/1/2 | — | 1–7 | M_2023_1_1 |

Všechny texty korpusu byly podrobeny automatizovanému syntaktickému rozboru. Za tímto účelem jsme využili modelu PDT-C 1.0, který byl speciálně vypracován pro český jazyk (Hajič a kol., 2020). Přestože autoři deklarují, že se snaží přidržovat se české skladebné terminologie tak, jak byla vypracována Šmilauerem a jak je standardně vyučována na českých základních a středních školách, zároveň upozorňují na fakt, že počítačové zpracování jazyka se od ní musí nutně odlišovat, už jen z toho důvodu, že je při něm potřeba anotovat (tj. syntakticky označkovat) každé grafické slovo, včetně interpunkce (srov. Hajič a kol., 2004). Mizí tak např. přísudky složené (infinitivy, které je tvoří, jsou anotovány jako předměty), naproti tomu se šířeji chápe atribut (jsou za něj považovány např. i složky adres, ty se vztahují k určitému podstatnému jménu a rozvíjejí jej, často jako identifikační nebo lokalizační údaj;

v analytické rovině se tak za atribut může považovat například jméno ulice, číslo domu či název města – všechny dané složky plní roli rozvíjejícího členu a jako takové jsou anotovány) a posuny se objevují též u dalších větných členů.

Jednotlivé rozdíly lze exemplifikovat na grafu uvedeném na obrázku 1. Graf zachycuje syntaktickou strukturu souvětí „Když jednou odběhla nakoupit, sundal jsem z háčku klíče a zkoušel dveře skrývající tajemství odemknout". Informace o větných členech (první řádek pod daným slovem) jsou doplněny rovněž morfologickou charakteristikou (druhý řádek), která je však pro nás v tuto chvíli bezpředmětná. Jak je patrné, model přiřazuje syntaktickou značku důsledně každé grafické pozici; např. spojky podřadicí („když") jsou součástí tzv. pomocných větných členů (zde konkrétně AuxC), podobně jako předložky (AuxP). Koordinační struktury jsou označeny jako „Coord"; toto rozhodnutí umožňuje analyzovat jako jeden větněčlenský typ např. spojku „a" a spojení asyndentické (bezespoječné), kde je jako „Coord" označena čárka (např. v souvětí „přišel, posadil se a dal si večeři" jsou tak jako „Coord" označena čárka a spojka „a"). Jak už bylo uvedeno výše, přísudky („sundal", „zkoušel") jsou důsledně rozkládány (pomocné sloveso „být" je označeno jako AuxV) a je u nich rovněž zohledněn fakt, že jsou ve větě souřadně spojené (jsou označeny jako Pred_Co). Jde-li o přísudek vedlejší věty („odběhla"), pak je tento určován jako ten větný člen, kterým je zastoupena celá klauze; v našem příkladu je tedy „odběhla" určeno jako příslovečné určení (adv), protože jde o predikát vedlejší věty příslovečné.

Na základě takto vygenerovaných větných členů navrhujeme pět indikátorů, pomocí nichž budeme měřit syntaktickou komplexitu textu. Naše úvahy vycházejí jednak z výše uvedených kvantitativněsyntaktických prací, jednak z prestižních českých stylistických zdrojů (srov. Hoffmannová a kol., 2016; Cvrček a kol., 2020). Jde o tyto markery (první tři normalizujeme jejich vztažením ke všem syntakticky anotovaným pozicím):

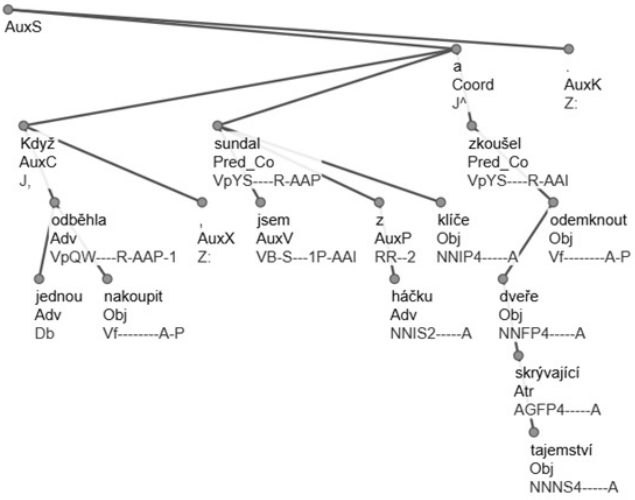
(1) rozšířená atributivita (atr_plus), která zahrnuje přívlastky („hezká dívka"; atr), přívlastky souřadně spojené („hezká a chytrá dívka"; atr_co), přívlastky, které jsou součástí apozic („postavení UK (Univerzity Karlovy)"; atr_ap), přívlastky, jež spoluutvářejí parenteze („její (neúspěšný) pokus"; atr_pa), doplňky („udělal to jako otec"; atv) a doplňky souřadně spojené („udělal to jako otec a manžel"; atv_co);

(2) frekvence příslovečných určení (adv), která zahrnují příslovečná určení („mluvil nezištně"; adv), příslovečná určení souřadně spojená („mluvil nezištně a příjemně"; adv_co), příslovečná určení v apozicích („odpoví běžným řešením: devalvací"; adv_ap) a příslovečná určení v parentezích („složil jí (na chodbě) poklonu"; adv_pa);

(3) míra apozice (apos), přičemž jako apozici chápeme uváděcí výraz, např. dvojtečku v již zmíněném spojení „odpoví běžným řešením: devalvací";

(4) frekvence jmenných částí přísudku (pnom), jimiž rozumíme jmenné části přísudku („byl nevrlý“; pnom), jmenné části přísudku souřadně spojené („byl nevrlý a podrážděný“; pnom_co) a jmenné části přísudku v apozici („bylo to enigmatické, tedy velmi záhadné“; pnom_ap);

(5) míra podřadnosti (subord), kterou vypočítáme jako podíl subordinačních pomocných větných členů (označených jako AuxC, viz obrázek 1) ku součtu AuxC a různých kombinací koordinačních grafických slov/znaků (coord – „přišel sám a hned“; coord_co – „přišel a posadil se, ale nebyl ve své kůži“; coord_ap – „stanovil symboly, a to znak a prapor“; coord_pa – „souhlasila (ale neusmívala se) s návrhem“).




Obrázek 1
*Příklad větného grafu vygenerovaného na základě modelu PDT-C 1.0.*

K těmto pěti indikátorům přidáváme dva další, které stojí na specifické syntaktické jednotce – lineárním dependenčním segmentu (LDS). Jeden LDS je tvořen slovy, která mají mezi sebou bezprostřední syntaktickou závislost, sousedí spolu v lineárním uspořádání věty a jsou součástí téže klauze (= věty v souvětí). Jakmile není jedna z těchto podmínek dodržena, vzniká LDS nový. Příkladem může být rozbor věty, kterou jsme uvedli výše („Když jednou odběhla nakoupit, sundal jsem z háčku klíče a zkoušel dveře skrývající tajemství odemknout.“). Mezi prvním

slovem („když“) a druhým („jednou“) není přímá syntaktická závislost, takže první LDS obsahuje pouze první slovo věty. Naproti tomu mezi slovem „jednou“ a „odběhla“ tato závislost existuje, a slova spolu navíc ve větě sousedí, takže LDS pokračuje dále; rovněž slova „odběhla“ a „nakoupit“ splňují všechny výše zmíněné podmínky, a proto budou součástí jednoho segmentu. Další slovo, které lineárně následuje, je již součástí další klauze, proto zde tento LDS končí – obsahuje tedy slova „jednou“, „odběhla“ a „nakoupit“. Replikací tohoto postupu bychom došli k rozkladu dané věty na tyto LDS:

[když], [jednou odběhla nakoupit], [sundal jsem], [z háčku], [klíče], [a zkoušel], [dveře skrývající tajemství], [odemknout].

Daná věta tedy obsahuje osm LDS. Je patrné, že vyšší počet LDS bude typický pro komplexnější texty, což se už v několika didakticky zaměřených výzkumech potvrdilo (srov. Nogolová a kol., 2023a; Nogolová a kol., 2023b). V našem bádání tuto jednotku zužitkujeme v podobě následujících dvou indexů:

(1) ACL, který značí průměrnou délku klauze (= věty v souvětí) v počtu LDS;
(2) ALDSL, jenž zachycuje průměrnou délku LDS v počtu slov.

Celkem tedy budeme u textů počítat sedm syntaktických indexů (atr_plus, adv, apos, pnom, subord, ACL a ALDSL); získáme tak pro každý index čtyři sady výsledků (E8, E6, E4 a M). Rozdíly mezi jednotlivými typy didaktických testů budeme statisticky testovat; využijeme Kruskalova–Wallisova testu, který se zaměřuje na zjišťování diferencí mezi třemi a více vzorky. Zároveň má tu výhodu, že je neparametrický, nepředpokládá tedy normální rozdělení dat (Wayne, 1990). Pokud vyjde výsledek tohoto testu jako statisticky signifikantní (= pravděpodobnost, že rozdíly mezi sadami textů jsou výsledkem náhody, bude menší než 5 %), budeme zkoumat, mezi jakými páry se tato signifikance vyskytuje, a to pomocí Dwassova–Steelova–Critchlowova–Flignerova testu (Critchlow – Fligner, 1991). Statistické analýzy bude provádět software NCSS (NCSS LLC, 2024).

Na závěr bude provedena analýza hlavních komponent, jež slouží ke snížení dimenzionality dat (v našem výzkumu pracujeme se sedmi indexy) a poukáže na to, zdali lze texty rozdělit do skupin (klastrů). Texty budou zachyceny v dvourozměrném prostoru, který vymezí dvě hlavní komponenty, zachycující určité procento variability v datech. Do analýzy vstoupí robustně standardizované výsledky výpočtů indexů (pomocí mediánu a mezikvartilového rozpětí). Prezentovány budou rovněž míry, s nimiž jednotlivé indexy přispívají na obě hlavní komponenty (tzv. loadingy). K provedení analýzy poslouží skript v Pythonu, přičemž využity budou balíčky pandas (McKinney, 2010), numpy (Harris a kol., 2020), matplotlib (Hunter, 2007)

a scikit-learn (Pedregosa a kol., 2011). K sestavení skriptu bylo využito ChatuGPT (verze 4o). Skript a plné výsledky výzkumu jsou volně k dispozici na platformě GitHub.[1]

## 3. Výsledky

Výsledky základní kvantitativněsyntaktické analýzy jsou prezentovány v tabulkách 3 a 4. Z první z nich vyplývá, že statisticky signifikantní rozdíly se mezi sadami didaktických textů objevují u rozšířené atributivity (atr_plus), frekvence příslovečných určení (adv), četnosti apozic (apos) a průměrné délky klauzí (ACL). Mírně nad přijatou mezí významnosti zůstává průměrná délka LDS (ALDSL). Co se týče konkrétních rozdílů mezi texty z hlediska párového testování, jako specifické se ve všech sledovaných případech ukazují maturitní testy; vykazují statisticky signifikantně vyšší míru rozšířené atributivity a nižší četnosti příslovečných určení oproti E6 a E8, vyšší frekvence apozic než všechny přijímací testy a delší průměrnou délku klauzí než E8. Je tedy možné je obecně považovat za komplexnější, neboť texty v nich obsažené preferují složité syntaktické struktury. Je poněkud překvapivé, že maturity se vyznačují nižším množstvím adverbialií, která by bylo možné očekávat např. u populárně-vědeckých textů; roli zde však může sehrávat odklon od narativního charakteru komunikátů, jenž je typický spíše pro přijímací zkoušky a který charakterizují důležitá příslovečná určení času a místa, jež čtenáři umožňují se v příběhu orientovat.

Tabulka 3
*Průměrné hodnoty a směrodatné odchylky syntaktických indikátorů pro jednotlivé sady testů. Sloupec nadepsaný „H" obsahuje testové statistiky Kruskalova–Wallisova textu. Ve sloupci následujícím je uveden jejich přepočet na p-hodnotu, který je v případě statické významnosti zvýrazněn tučně (jako hladina významnosti bylo stanoveno 0,05): * = p < 0,05, ** = p < 0,01, *** = p < 0,001.*

| | E8 | E8 [SD] | E6 | E6 [SD] | E4 | E4 [SD] |
|---|---|---|---|---|---|---|
| atr_plus | 0,1755 | 0,0720 | 0,1821 | 0,0720 | 0,1928 | 0,0774 |
| adv | 0,1387 | 0,0276 | 0,1410 | 0,0270 | 0,1345 | 0,0303 |
| apos | 0,0029 | 0,0041 | 0,0032 | 0,0041 | 0,0031 | 0,0043 |
| pnom | 0,0153 | 0,0093 | 0,0162 | 0,0108 | 0,0145 | 0,0080 |

[1] https://github.com/KCJFFOU/syntax_didakticke_studie.

| | | | | | | |
|---|---|---|---|---|---|---|
| subord | 0,3027 | 0,1452 | 0,3040 | 0,1486 | 0,3083 | 0,1550 |
| ACL | 3,6653 | 0,8535 | 3,8689 | 0,9408 | 3,9259 | 0,9318 |
| ALDSL | 1,7602 | 0,0822 | 1,7644 | 0,0744 | 1,7637 | 0,0840 |
| | M | M [SD] | H | p-value | | |
| atr_plus | 0,2223 | 0,0838 | 15,0395 | **0,0018**** | | |
| adv | 0,1230 | 0,0357 | 13,2883 | **0,0041**** | | |
| apos | 0,0052 | 0,0052 | 10,6475 | **0,0138*** | | |
| pnom | 0,0152 | 0,0093 | 0,7418 | 0,8633 | | |
| subord | 0,2835 | 0,1742 | 0,7369 | 0,8645 | | |
| ACL | 4,0604 | 0,8637 | 8,9441 | **0,03*** | | |
| ALDSL | 1,7893 | 0,0844 | 6,8582 | 0,0766 | | |

Tabulka 4

*Výsledky statistického testování v párech. Pracujeme pouze s indexy, které byly vyhodnoceny jako nositelé statisticky signifikantních diferencí mezi sadami (viz tabulku 3). Sloupec „W*" prezentuje statistiky Dwassova–Steelova–Critchlowova–Flignerova testu. Ve sloupci následujícím je uveden jejich přepočet na p-hodnotu, který je v případě statické významnosti zvýrazněn tučně (jako hladina významnosti bylo stanoveno 0,05): * = p < 0,05; ** = p < 0,01; *** = p < 0,001.*

| | E8 vs E6 | | E8 vs E4 | | E8 vs M | |
|---|---|---|---|---|---|---|
| | W* | p-hodnota | W* | p-hodnota | W* | p-hodnota |
| atr_plus | 1,0132 | 0,8906 | 2,1082 | 0,4431 | 5,0567 | **0,002**** |
| adv | 0,6407 | 0,9691 | 1,447 | 0,7359 | 4,1231 | **0,0186*** |
| apos | 0,5595 | 0,979 | 0,2507 | 0,998 | 4,2965 | **0,0127*** |
| ACL | 1,8901 | 0,5396 | 2,69 | 0,2271 | 4,1353 | **0,0182*** |
| | E6 vs E4 | | E6 vs M | | E4 vs M | |
| | W* | p-hodnota | W* | p-hodnota | W* | p-hodnota |
| atr_plus | 1,2957 | 0,7962 | 4,4085 | **0,0099**** | 3,1467 | 0,1165 |
| adv | 2,2727 | 0,3746 | 4,723 | **0,0047**** | 2,7615 | 0,2062 |
| apos | 0,2656 | 0,9977 | 3,7646 | **0,0389*** | 4,0002 | **0,0242*** |
| ACL | 0,6901 | 0,9618 | 2,3892 | 0,3292 | 1,5092 | 0,7096 |

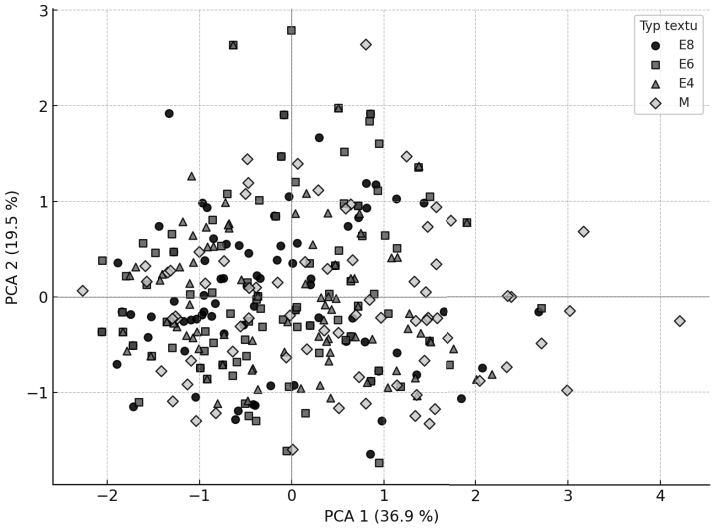
Negativní vztah mezi rozšířenou atributivitou a četností příslovečných určení je možné vypozorovat také z výsledků analýzy hlavních komponent (viz tabulku 5). První komponentu, kterou lze považovat za stěžejní (vysvětluje 37 % variability dat), je možné interpretovat pomocí kontrastu nominalita/verbalita. Texty s vysokými hodnotami na této komponentě jsou atributivně bohaté, užívají apozice, dlouhé klauze (měřeno v počtu LDS) a dlouhé LDS; naproti tomu ty texty, které v této komponentě nabývají nízkých hodnot, se vyznačují vysokým podílem příslovečných určení a subordinace. Zatímco pro verbálně založené vyjadřování je typické rozvíjení slovesné fráze (které se většinou děje prostřednictvím příslovečných určení), nominálnost charakterizuje tendence ke složitým frázím jmenným, jež lze vystavět na základě atributů. Rozvětvené nominální fráze pak velmi výrazně souvisejí s větším množstvím LDS v klauzích, protože např. kumulace přívlastků shodných generuje vyšší počet lineárních dependenčních segmentů (např. ve frázi „naše nová jarní slevová akce" jsou čtyři). Druhá komponenta je naproti tomu velmi málo výrazná; charakterizuje ji pouze vysoký podíl přísudků jmenných se sponou. Celkem vysvětluje 20 % variability dat; první dvě komponenty tedy vysvětlují v úhrnu 57 % variability.

Tabulka 5
*Loadingy jednotlivých indexů na prvních dvou komponentách PCA.*

| index | PCA1 | PCA2 |
|---|---|---|
| ACL | 0,43 | 0,01 |
| ALDSL | 0,41 | -0,20 |
| atr_plus | 0,52 | -0,07 |
| adv | -0,50 | -0,01 |
| apos | 0,45 | -0,05 |
| pnom | 0,20 | 0,78 |
| subord | -0,37 | 0,06 |

Vizualizované výsledky analýzy hlavních komponent (viz obrázek 2) indikují absenci zřetelně odlišitelných klastrů ve zkoumaném vzorku; nedochází tedy k jednoznačnému rozdělení textů do skupin, případně k vydělení zástupců některého ze zkoumaných didaktických textů. V oblasti vysokých hodnot první hlavní komponenty se objevují většinově texty z maturitních didaktických testů, ale mnoho z nich je rozptýleno i na jiných místech grafu. Absenci výraznějších rozdílů mezi testy přisuzujeme žánrové diverzitě zkoumaných textů – v korpusu se

objevují zhuštěné medailonky autorů či prezentace uměleckých směrů, úryvky z beletrie, publicistika, popularizační články i korespondence. Tento rozptyl má za následek značnou variabilitu ve výsledcích, která znemožňuje rozdělení textů do vymezených skupin.




Obrázek 2
*Rozložení textů v dvoudimenzionálním prostoru vymezeném prvními dvěma hlavními komponentami.*

## 4. Závěry

Syntaktická analýza zjistila ve zkoumaném vzorku didaktických testů statisticky signifikantní rozdíly, co se týče rozšířené atributivity (atr_plus), frekvence příslovečných určení (adv), četnosti apozic (apos) a průměrné délky klauze počítané v lineárních dependenčních segmentech (ACL). Při podrobnějším zkoumání se ukázalo, že tyto rozdíly se objevují výhradně mezi maturitním didaktickým testem a testy přijímacích zkoušek; v žádném indikátoru se nelišily přijímací testy mezi sebou. Maturitní testy se tak vyznačují vyšším podílem atributů a nižšími frekvencemi příslovečných určení oproti přijímacím zkouškám do osmiletého a šestiletého středoškolského studia; ve srovnání se všemi přijímacími testy užívají více apozic

a vůči přijímacím testům do osmiletého studia se vymezují v průměru delšími klauzemi v počtu LDS. Obecně tak lze shrnout, že maturitní texty kladou důraz na komplexní větnou stavbu, kterou charakterizují komplikované dependenční vztahy a bohaté nominální fráze.

S tím, že rozdíly jsou omezeny jen na určité indexy a dvojice, souvisí i výsledky analýzy hlavních komponent; ty ukázaly, že texty nelze jednoznačně rozklasifikovat do klastrů, tvoří spíše kontinuum. První komponenta je organizuje především podle preference nominálního a verbálního vyjadřování a syntaktické složitosti. Roli ve výsledcích může sehrávat, jak už bylo uvedeno, především žánrová rozrůzněnost vzorku.

Představený výzkum je v několika směrech limitovaný. Především je třeba brát v potaz chybovost syntaktické anotace a fakt, že v případě analýzy hlavních komponent uvažujeme pouze první dvě (i když množství vysvětlené datové variability je poměrně vysoké). Další omezení spočívá ve výběru indikátorů, který byl založen na autoritativních zdrojích z české stylistiky, ale vždy v sobě obsahuje určitou míru subjektivity. Také je třeba zmínit časový rámec korpusu, který pokrývá posledních pět let, což nemusí být úsek dostatečný k tomu, aby bylo možné hovořit o dlouhodobých tendencích.

Přes uvedená omezení pokládáme zjištěné skutečnosti za velmi relevantní pro českou didaktiku; jsou s to poskytnout lingvistická vodítka pro měření syntaktického aspektu komplexity textu a mohou přispět do již zmíněné diskuse nad koncepcí didaktických testů. Hlavní aplikační potenciál výzkumu vidíme v testové diagnostice, kdy bude možné na základě komplexních lingvistických dat určit, zdali se v daném didaktickém testu neobjevuje zvýšené množství textů jazykově náročných. Zároveň se výsledky dají využít i při tvorbě testů – např. při posuzování, zda je daný text vhodný pro určitou přijímací zkoušku. K tomu je však třeba lingvistický výzkum zkompletovat, především analýzou textových sekvencí v jednotlivých testech a kvantitativněLingvistickým rozborem zaměřeným na lexikální indexy (např. hustotu a diverzitu).

**Poděkování**

Výzkum byl podpořen Filozofickou fakultou Ostravské univerzity (projekt: SGS13/FF/2024 *Čtenářská gramotnost pohledem současných lingvistických metod*) a Moravskoslezským krajem (v rámci programu Podpora vědy a výzkumu v MSK 2023 v projektu *Podpora talentovaných studentů doktorského studia na Ostravské univerzitě VII.*, ev. č. smlouvy 00516/2024/RRC).

Autoři děkují oběma anonymním recenzentům za pečlivé a kritické posouzení rukopisu, zejména za statistické postřehy. Metodologické připomínky významně přispěly ke zkvalitnění textu a byly implementovány do přepracované verze.

## Literatura


Baumgartnerová, G. (2021–2022): Několik slov ke koncepci didaktických testů z českého jazyka a literatury. *Český jazyk a literatura*, roč. 72, č. 3, s. 126–133.

Cermat. (2019): Jednotná přijímací zkouška. *Centrum pro zjišťování výsledků vzdělávání.* https://prijimacky.cermat.cz/

Cermat. (2019): Maturitní zkouška. *Centrum pro zjišťování výsledků vzdělávání.* https://maturita.cermat.cz/

Critchlow, D. E. – Fligner, M. A. (1991): On distribution-free multiple comparisons in the one-way analysis of variance. *Communications in Statistics – Theory and Methods*, sv. 20, č. 1, s. 127–139.

Cvrček, V. – Laubeová, Z. – Lukeš, D. – Poukarová, P. – Řehořková, A. – Zasina, A. J. (2020): *Registry v češtině.* Praha: Nakladatelství Lidové noviny.

Čaban, M. (2024, říjen 24): *Komentář: Končí šéf Cermatu, autor digitalizace, která se povedla. Až moc*. Seznam Zprávy. https://www.seznamzpravy.cz/clanek/domaci-zivot-v-cesku-komentar-konci-sef-cermatu-autor-digitalizace-ktera-se-povedla-az-moc-263144

Čech, R. – Mačutek, J. – Kubát, M. – Koščová, M. (2022): Does an author leave a syntactic footprint? In: Misuraca, M. – Scepi, G. – Spano, M. (Eds.): *Proceedings of the 16th International Conference on statistical analysis of textual data* (s. 221–228). Napoli: Vadistat Press.

Davidson, D. (2004): *Subjektivita, intersubjektivita, objektivita*. Praha: Filosofia.

Graesser, A. C. – McNamara, D. S. – Kulikowich, J. M. (2011): Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, sv. 40, č. 5, s. 223–234.

Hajič, J. – Bejček, E. – Hlavacova, J. – Mikulová, M. – Straka, M. – Štěpánek, J. – Štěpánková, B. (2020): Prague Dependency Treebank – Consolidated 1.0. In: Calzolari, N. – Béchet, F. – Blache, P. – Choukri, K. – Cieri, C. – Declerck, T. – Goggi, S. – Isahara, H. – Maegaard, B. – Mariani, J. – Mazo, H. – Moreno, A. – Odijk, J. – Piperidis, S. (Eds.): *Proceedings of the Twelfth Language Resources and Evaluation Conference* (s. 5208–5218). Marseille: European Language Resources Association.

Hajič, J. – Panevová, J. – Buráňová, E. – Urešová, Z. – Bémová, A. – Kárník, J. – Štěpánek, J. – Pajas, P. (2004): *Anotace na analytické rovině: Návod pro anotátory. Technická zpráva TR-2004-23*. Praha: ÚFAL MFF UK.

Harris, C. R. – Millman, K. J. – van der Walt, S. J. – Gommers, R. – Virtanen, P. – Cournapeau, D. – Wieser, E. – Taylor, J. – Berg, S. – Smith, N. J. – Kern, R. – Picus, M. – Hoyer, S. – van Kerkwijk, M. H. – Brett, M. – Haldane, A. – Fernández del Río, J. – Wiebe, M. – Peterson, P. – Gérard-Marchant, P. – Sheppard, K. – Reddy,

T. – Weckesser, W. – Abbasi, H. – Gohlke, C. – Oliphant, T. E. (2020): Array programming with NumPy. *Nature*, sv. 585, č. 7825, s. 357–362.
Hoffmannová, J. – Homoláč, J. – Chvalovská, E. – Jílková, L. – Kaderka, P. – Mareš, P. – Mrázková, K. (2016): *Stylistika mluvené a psané češtiny*. Praha: Academia.
Hunter, J. D. (2007): Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, sv. 9, č. 3, s. 90–95.
Kabátová, Š. (2023, květen 26): *Glosa: Znovu a lépe. České školství je zralé na reparát*. Seznam Zprávy. https://www.seznamzpravy.cz/clanek/audio-podcast-dve-minuty-glosa-znovu-a-lepe-ceske-skolstvi-je-zrale-na-reparat-231576
Kubát, M. – Čech, R. – Chen, X. (2021): Attributivity and Subjectivity in Contemporary Written Czech. In: Chen, X. – Čech, R. (Eds.): *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)* (s. 58–64). Stroudsburg: Association for Computational Linguistics.
Kubát, M. – Čech, R. (2024): Attributivity and Subjectivity in Czech Journalism and Scientific Literature. *Bohemistyka*, roč. 24, č. 1, s. 3–14.
Kubát, M. – Mačutek, J. – Čech, R. – Nogolová, M. (2024): Automatic Genre Classification of Czech Texts Based on Syntactic Functions. In: Giordano, G. – Misuraca, M. (Eds.): *New Frontiers in Textual Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization* (s. 163–172). Cham: Springer.
Liptáková, Ľ. – Cibáková, D. (2013): *Edukačný model rozvíjania porozumenia textu v primárnom vzdelávaní*. Prešov: Prešovská univerzita, Pedagogická fakulta.
McKinney, W. (2010): Data Structures for Statistical Computing in Python. In: van der Walt, S. – Fernández, J. – Varroquaux, N. (Eds.): *Proceedings of the 9th Python in Science Conference* (s. 51–56). Austin: SciPy.
Místecký, M. – Radková, L. (2020): School and Gender in Numbers: A Stylometric Insight into the Lexis of Teenagers' Description Essays. *Glottometrics*, sv. 49, s. 52–65.
Místecký, M. – Radková, L. (2021–2022): Aktivní v popisu: Poznámky ke stylu žákovských slohových prací. *Český jazyk a literatura*, roč. 72, č. 3, s. 120–126.
Místecký, M. – Radková, L. – Stiborská, Ž. – Hrubá, D. (2025). Kvantitativní analýza morfologických charakteristik textů z didaktických testů posledních let. *Korpus – gramatika – axiologie*, č. 31, s. 25–37.
Münich, D. (2024, květen 23): *Komentář: Reforma přijímaček neskončila. Je třeba ji dotáhnout*. Seznam Zprávy. https://www.seznamzpravy.cz/clanek/domaci-zivot-v-cesku-komentar-reforma-prijimacek-neskoncila-je-treba-ji-dotahnout-252376
NCSS LLC. (2024): *NCSS 2024* (verze 24.0.3) [software]. https://www.ncss.com/download/ncss/latest/

Nogolová, M. – Čech, R. – Hanušková, M. – Kubát, M. (2023a): The Development of Sentence and Clause Lengths in Czech L2 Texts. *Korpus – gramatika – axiologie*, č. 28, s. 22–37.
Nogolová, M. – Hanušková, M. – Kubát, M. – Čech, R. (2023b): Linear Dependency Segments in Foreign Language Acquisition: Syntactic Complexity Analysis in Czech Learners' Text. *Jazykovedný časopis*, roč. 74, č. 1, s. 193–203.
OpenAI. (2024): *ChatGPT: Language model for generating textbased outputs* (verze 4o). https://openai.com
Pedregosa, F. ‐ Varoquaux, G. ‐ Gramfort, A. ‐ Michel, V. ‐ Thirion, B. ‐ Grisel, O. ‐ Blondel, M. ‐ Prettenhofer, P. ‐ Weiss, R. ‐ Dubourg, V. ‐ Vanderplas, J. ‐ Passos, A. ‐ Cournapeau, D. ‐ Brucher, M. ‐ Perrot, M. ‐ Duchesnay, É. (2011): Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, sv. 12, s. 2825–2830.
Pluskal, M. (1996): Zdokonalení metody pro měření obtížnosti didaktických textů. *Pedagogika*, roč. 46, č. 1, s. 62–76.
Průcha, J. (1998): *Učebnice: Teorie a analýzy edukačního média*. Brno: Paido.
Rafatbakhsh, E. ‐ Ahmadi, A. (2023): Predicting the difficulty of EFL reading comprehension tests based on linguistic indices. *Asian-Pacific Journal of Second and Foreign Language Education 8*, Article 40. https://doi.org/10.1186/s40862-023-00214-4
Štěpáník, S. (2018): Vliv nové podoby maturitní zkoušky z českého jazyka a literatury na vyučování ve výpovědích učitelů. *Pedagogická orientace*, roč. 28, č. 3, s. 435–471.
Štěpáník, S. (2020): *Výuka češtiny mezi tradicí a inovací*. Praha: Academia.
Wayne, D. W. (1990): *Applied Nonparametric Statistics*. Boston: PWS-KENT.

**Mgr. et Mgr. Michal Místecký, Ph.D.**
Katedra českého jazyka, Filozofická fakulta, Ostravská univerzita
*michal.mistecky@osu.cz*
ORCID: 0000-0002-9183-4435

**PhDr. Lucie Radková, Ph.D.**
Katedra českého jazyka, Filozofická fakulta, Ostravská univerzita
*lucie.radkova@osu.cz*
ORCID: 0000-0001-5193-4610