

Slovenská Wikipédia ako zdroj dát pre korpus chýb

RADOVAN GARABÍK

Jazykovedný ústav Ľ. Štúra SAV, v. v. i.

Abstract: The article describes an error corpus of Slovak, compiled from Wikipedia's revision history. The errors were automatically extracted and classified into spelling, typographical, lexicosemantic, diacritical, capitalization, or punctuation ones. The corpus is available via the NoSketch Engine corpus manager interface, presented in a parallel text format, and the classified errors are publicly available in a form of datasets. The corpus is a valuable resource documenting naturally occurring linguistic errors.

Keywords: corpus, error corpus, wikipedia, Slovak

Úvod

Článok je ďalším zo série opisujúcej významné korpusy Jazykovedného ústavu Ľ. Štúra Slovenskej akadémie vied. Predstavíme v ňom chybový korpus CHIBY, spracovávajúci revízie slovenskej Wikipédie^{1,2}.

Chybové korpusy (alebo korpusy chýb) sú špeciálnym typom korpusov, v ktorých sú v texte označené chyby, často aj s ich prípadným správnym variantom. Tieto korpusy sa v korpusovej lingvistike nevyskytujú často a nepatria medzi vlajkové lode veľkých korpusov. Napriek tomu majú neodškriepiteľný význam hlavne v oblasti trénovania nástrojov na automatickú opravu chýb v texte alebo pri výskume zameranom na osvojovanie si jazyka (či už ako L1, alebo ako cudzieho). Jednou z hlavných príčin menšieho rozšírenia chybových korpusov je ich celková náročnosť, pretože chyby je (často) nevyhnutné značkovať ručne.

Pri tvorbe chybových korpusov sa mnohokrát využívajú práce žiakov či študentov, ktoré sú obvykle tak či tak kontrolované a opravované, pretože cieľená ručná anotácia chybových korpusov by bola značne náročná. Podobne vznikol aj chybový korpus češtiny (Pala et al., 2003). Naopak, slovenský kor-

¹ <https://sk.wikipedia.org/>

² Článok si nekladie za cieľ rozhodnúť, či je Wikipédia/wikipédia proprium alebo apelatívum a podsúvať tak čitateľom dojem, že existuje jediná možná kapitalizácia tohto slova.

pus textov študentov učiacich sa slovenčinu ako cudzí jazyk ERRKORP sa vykonával touto ťažšou cestou cielenej anotácie chýb, čo umožnilo ich veľmi podrobnú a detailnú klasifikáciu (Gajdošová et al., 2024).

O Wikipédii

Wikipédiu netreba špeciálne predstavovať – ide o kolaboratívnu encyklopédiu, ktorá sa snaží pokryť celú oblasť ľudských vedomostí. Táto encyklopédia existuje vo veľkom množstve jazykových variantov vrátane slovenčiny, hoci slovenská verzia Wikipédie je pomerne slabšia a menšia než niektoré mutácie „väčších“ jazykov.

Z praktického pohľadu je veľmi významná aj licenčná politika Wikipédie, ktorá umožňuje zdieľať texty a ich odvodeniny pod pomerne atraktívnou licenciou Creative Commons Attribution/Share-Alike License 4.0 Unported (CC BY-SA 4.0).

Jednou z lákavých črt Wikipédie, na ktorej bol pôvodne založený raketový nárast jej popularity, je možnosť kolaboratívnej editácie (pôvodne bez obmedzenia, nutnosti registrácie a pod.) a následné vytvorenie pomerne silnej komunity redaktorov. Je charakteristická tým, že jednotlivé články sa veľmi často menia, pričom väčšina zmien predstavuje drobné úpravy a doplnky. Veľkú časť týchto zmien tvoria opravy chýb, ktoré nevedomky vniesol do pôvodného textu jeho autor. Wikipédia je preto veľmi cenným zdrojom práve pri vytváraní takéhoto chybového korpusu.

Na druhej strane, tento otvorený prístup má za následok jeden veľký problém, ktorý sa prejavuje aj pri tvorbe chybového korpusu, a to sú takzvaní vandali. V slangu redaktorov Wikipédie ide o prispievateľov, ktorých cieľom je buď anonymne sa zviditeľniť, alebo priamo škodiť. Snažia sa vnášať rôzne chyby, vyslovené klamstvá či nepravdy, často aj vulgárneho charakteru (na rozdiel od moderného trendu šírenia rôznych dezinformácií, ktoré majú buď skrytú agendu, alebo ich autori žijú vo svojom vlastnom svete a dezinformáciám naozaj veria).

Vo väčších wikipédiách, napríklad v anglickej, pôvodní redaktori článkov tieto chyby prakticky okamžite opravila, aspoň v prípade väčších, významnejších článkov. To však celkom neplatí pre slovenskú Wikipédiu, kde takéto vandalské zmeny môžu často zostať nepovšimnuté dlhé roky alebo aj navždy. Špecifickým problémom pri tvorbe chybového korpusu z dát Wikipédie je to,

že takéto zásahy nemožno považovať za opravy chýb, často ide o pravý opak (vnášanie chýb). Na druhej strane ani následná oprava vandalských zmien nie je opravou typických chýb v texte, a preto nie je vhodná pre chybový korpus.

Chybový korpus CHIBY

Náš chybový korpus CHIBY obsahuje revízie textov (na úrovni odsekov) s automatickou anotáciou typu zmien; cieľom je maximalizovať inklúziu, takže zahrnuté sú (po uplatnení jednoduchých filtrov na vhodnosť a legitimitu zmien) všetky za sebou nasledujúce revízie, čo, bohužiaľ, zvyšuje pravdepodobnosť, že do korpusu cez filtre prenikli aj vandalské zmeny.

Klasifikácia chýb v našom korpuse je inšpirovaná klasifikáciou opísanou J. Buštom a kol. (2009). Rozlišujeme nasledujúce anotácie zmien (zodpovedajú štruktúram v korpusovom manažéri, kde sa zobrazujú intuitívnymi skratkami):

- spelling – zmena v pravopise slova
- punct – zmena v interpunkcii
- typographical – typografická úprava
- diacritics – zmena (iba) v diakritike alebo zmena zachováajúca ekvivalenciu (po NFKD dekompozícii) textu v Unicode (UAX15)
- capitalisation – zmena (iba) vo veľkosti písmen
- lexicosemantic – lexikálno-sémantická, všeobecná zmena slova alebo viacerých slov
- unclassified – rôzne iné, nezaradené zmeny

Revízie v bielych znakoch nie sú osobitne označené, prejavajú sa v rozdielnych medzerách v paralelnom zobrazení textu.

Z uvedenej klasifikácie je zrejmé, že typy chýb sa niekedy prekrývajú. Zmenu v diakritike by sme mohli klasifikovať aj ako zmenu v pravopise slova alebo zmena v pravopise môže byť chápaná aj ako lexikálno-sémantická, najmä ak ide o zmenu z jedného existujúceho slova na iné.

Štruktúra korpusu

Na rozdiel od väčšiny slovenských písaných korpusov nie je tento korpus lematizovaný, keďže jeho ťažisko je práve v „zle napísaných“ slovách, kde by sa nedala dosiahnuť želanú presnosť lematizácie. Slová majú iba

dva atribúty, *word* je slovo v tvare, v ktorom sa vyskytuje v texte (vrátane zachovania veľkosti písmen), *lc* je slovo, v ktorom sú všetky písmená konvertované z veľkých na malé.

Korpus je členený do dokumentov, každý dokument zodpovedá jednej revízii jedného článku vo Wikipédii, kde zachovávame iba tie vety, v ktorých bola uskutočnená zmena. Dokumenty sa ďalej delia na vety <s> (ale nie na odseky), podobne ako v ostatných korpusoch JÚLŠ¹. V tabuľke 1 uvádzame všetky štruktúry korpusu.

Tabuľka 1 – Štruktúry korpusu

doc	dokument (článok Wikipédie)
doc.page	názov článku
doc.contributor	autor zmeny
doc.comment	opis zmeny (komentár zadávaný autorom zmeny)
doc.date	dátum zmeny (vo formáte ISO 8601)
doc.wordcount	počet slov v dokumente
s	veta
g	na tomto mieste sa medzi tokenmi nenachádza medzera
lexicosemantic punct unclassified diacritics spelling capitalisation typographical	indikuje miesto, na ktorom bol zmenený text, a typ zmeny

Predchádzajúce korpusy spracovávajúce editovacie zmeny vo Wikipédii

Zrejme prvým cielene lingvistickým korpusom založeným na opravách Wikipédie bol francúzskojazyčný chybový korpus (Max – Wisniewski, 2010). Poľský korpus je zdokumentovaný u Grundkiewicza (2013), použitý softvér autori nazvali Wiki Edits (Grundkiewicz – Junczys-Dowmunt, 2014) a okrem poľského ho použili aj na vytvorenie anglického chybového korpusu.

¹ <https://www.juls.savba.sk/tools.html#Korpusy>

Ďalej existuje málo zdokumentovaný softvér na extrakciu revízií pochádzajúci z Ústavu formálnej a aplikovanej lingvistiky, Matematicko-fyzikálnej fakulty Univerzity Karlovej².

Bakalárska práca (Kletečka, 2017) tiež opisuje tvorbu chybového korpusu (češtiny) na zmenách vo Wikipédii. Použitý softvér je tiež dostupný ako Open Source³.

Tvorba korpusu

Po analýze dostupných možností (dostupné nástroje uvedené v predchádzajúcej časti poväčšine nefungovali na novšej verzii Wikipédie alebo mali rôzne iné nedostatky a problémy) sme sprevádzkovali a upravili Wiki Edits. Tento softvér používa heuristiku na zistenie vandalských zmien – na základe kľúčových slov v komentári nasledujúcej zmeny, ktoré indikujú vrátenie vandalskej verzie do predchádzajúceho stavu. Adaptácia softvéru na iný jazyk spočíva hlavne v zmene tejto funkcie, väčšinou poskytnutím zoznamu takýchto kľúčových slov. Podobne sme postupovali aj my, tieto kľúčové slová sme získali z frekvenčnej analýzy komentárov zodpovedajúcich zmenám slovenskej Wikipédie. Na informáciu uvádzame regulárny výraz na zistenie vandalských zmien v originálnej poľskej verzii:

„`anulowan|wycofan|cofnie|cofnie|przywróc|przywróc|revert|rewert`“
a v našej slovenskej verzii: „`revert|vráten|vraten|vandal|obnov|revízia|revízia`“.

Wiki Edits sme neskôr upravili, aby fungoval aj v súčasnej (v čase písania tohto článku) verzii⁴ programovacieho jazyka Python 3, táto verzia Wiki Edits je dostupná na adrese <https://github.com/garabik/wikiedits3>.

Tabuľka 2 – najčastejšie zmeny v kapitalizácii

<i>počet zmien</i>	<i>originál</i>	<i>revízia</i>
109	v	V
100	Druhej	druhej
95	prof	Prof
83	The	the
77	V	v
72	Majstrovstvách	majstrovstvách

² <https://github.com/ufal/wiki-error-corpus>

³ <https://github.com/jirkle/ErrCorp>

⁴ <https://www.python.org/>

Tabuľka 3 – najčastejšie zmeny v diakritike

<i>počet zmien</i>	<i>originál</i>	<i>revízia</i>
164	svojim	svojím
143	kedže	keďže
137	su	sú
134	prevážne	prevažne
114	km2	km²
109	Kedže	Keďže

Tabuľka 4 – najčastejšie lexikálno-sémantické zmeny

<i>počet zmien</i>	<i>originál</i>	<i>revízia</i>
2369		približne
1600	je	bol
1457	sa	
1254		sa
1099	avšak	ale
1009		v

Poznámka: Prázdna bunka znamená, že išlo buď o doplnenie (originál), alebo vymazanie (revíziu) slova.

Tabuľka 5 – najčastejšie zmeny v pravopise

<i>počet zmien</i>	<i>originál</i>	<i>revízia</i>
904	se	sa
537	súčastnosti	súčasnosti
298	ktorý	ktorí
278	z	s
269	s	z
210	v	vo

Zhrnutie

Korpus CHIBY poskytuje verejnosti nový, významný zdroj lingvistic-
kých dát slovenského jazyka. Korpus je prístupný na adrese <https://www.juls.savba.sk/errcorp.html> prostredníctvom korpusového manažéra NoSketch Engine, vo formáte paralelného korpusu (originál – revízia). Celkovo korpus

obsahuje 952 933 záznamov revízií (po odstránení vandalských zmien). Pre zaujímavosť tu uvádzame v tabuľkách 2 až 5 najčastejšie zmeny pre rôzne kategórie revízií. Datasets kompletných tabuliek sú dostupné na stránke korpusu. Korpus chýb sa občas aktualizuje, zachytáva tak pomerne aktuálne nové verzie slovenskej Wikipédie vrátane aktuálnych alebo novších spoločenských jazykových prejavov.

LITERATÚRA

BUŠTA, Jan – JAKUBÍČEK, Miloš – HLAVÁČKOVÁ, Dana – PALA, Karel: Classification of Errors in Text. In RASLAN 2009 : Recent Advances in Slavonic Natural Language Processing. 1. vyd. Brno: Masaryk University, 2009. s. 109 – 119. ISBN 978-80-210-5048-8.

GAJDOŠOVÁ, Katarína – LEVICKÁ, Jana – UHLIARIK, Ivor: Korpus textov študentov učiacich sa slovenčinu ako cudzí jazyk. In: Jazykové chyby v slovenčine ako cudzom jazyku na báze akvizičného korpusu. Ed. Jana Pekarovičová. Bratislava: Univerzita Komenského v Bratislave 2024, s. 30 – 73. ISBN 978-80-223-5868-2.

GRUNDKIEWICZ, Roman: Automatic Extraction of Polish Language Errors from Text Edition History. In: Habernal, I., Matoušek, V. (eds) Text, Speech, and Dialogue 2013. Lecture Notes in Computer Science, zv. 8082. Springer, Berlin, Heidelberg 2013, s. 129 – 136. https://doi.org/10.1007/978-3-642-40585-3_17.

GRUNDKIEWICZ, Roman – JUNCZYS-DOWMUNT, Marcin: The WikEd Error Corpus: A Corpus of Corrective Wikipedia Edits and Its Application to Grammatical Error Correction. In: Przepiórkowski, A., Ogrodniczuk, M. (eds) Advances in Natural Language Processing. NLP 2014. Lecture Notes in Computer Science, zv. 8686. Cham: Springer 2014, s. 478 – 490. https://doi.org/10.1007/978-3-319-10888-9_47.

KLETEČKA, Jiří.: Wikipedia Learner's Corpus. Bakalárska práca. Brno: Fakulta informatiky, Masarykova Univerzita 2017.

MAX, Aurélien – WISNIEWSKI, Guillaume: Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation 2010. S. 3143 – 3148.

PALA, Karel – RYCHLÝ, Pavel – SMRŽ, Pavel: Text corpus with errors. In: Text, Speech and Dialogue: Sixth International Conference, TSD 2003, Berlin, Springer Verlag 2003, s. 90 – 97.

CC BY-SA 4.0. Creative Commons Attribution/Share-Alike License 4.0 Unported.

<https://creativecommons.org/licenses/by-sa/4.0/> [cit. 27. 10. 2025].

UAX15. Unicode Standard Annex #15: Unicode Normalization Forms Najnovšia verzia:

<https://www.unicode.org/reports/tr15/> [cit. 27. 10. 2025].