Základy odborného stylu

K. Pala

pala@fi.muni.cz

Centrum zpracování přirozeného jazyka FI MU

21.2.2019

O čem bude řeč?

- Komunikace v přirozeném jazyce (PJ)
- Zvuky a znaky
- Pravopisné systémy
- Pravidla (českého) pravopisu
- Jazyková správnost
- Pravopisné chyby a jejich opravování
- Korektory – pravopisné, gramatické
- Komunikace H-C, jednocestná: dvoucestná
- Počítače a PJ, software pro práci s PJ, vztah k UI
- Poznámka k předmětu ZOS (VB000)

Pozitivní komunikace

- Výchozí axiom: nelze nekomunikovat
- Naše civilizace stojí na komunikaci – je to základ pro vědu, techniku, kulturu, pro veškeré znalosti lidstva
- Musí splňovat jisté standardy, aby byla efektivní
- Pro úspěšnou komunikaci potřebujeme vhodná pravidla, normy, což se často podceňuje
- Většina lidské komunikace probíhá v přirozeném jazyce
- Má dvě základní podoby: mluvenou a psanou
- Mluvený jazyk je výchozí, znalosti ukládáme v psané f.
- Potřeba norem pro přechod od mluvené podoby k psané

Mluvený a psaný jazyk

- Mluvený jazyk je primární (cca 5000 jazyků světa)
- Psaný jazyk je až sekundární – je to ale paměť lidstva
- Psané texty v přirozeném jazyce vyžadují přiřazení zvuků jazyka (hlásek) psaným znakům (písmenům)
- Úplná reprezentace zvuků v jednotlivých jazycích světa, k tomu účelu existuje: International Phonetic Alphabet (IPA)
- Reprezentace zvuků: různé pravopisné systémy
- Jsou základem veškeré civilizace a kultury obecně
- Abecední písma – latinka, cyrilice a např. mchedruli (gruz.)
- Ideografická písma (logografická) – čínština, cca 50 tis. znaků (základní soubor čítá 9-12 tis. znaků, standard)
- Minimální slovník 1200 slov v ang., školní standard 6-8000
- Slabičná písma – japonština (hiragana, katakana, kandži), korejština (hangul), jiná – např. amharština (ge‘ez)

Přiřazení zvuků znakům

- Kolik hlásek (fonémů) má čeština?
- Kolik písmen (znaků) má čeština?
- Kolik hlásek (fonémů) je v angličtině?
- Kolik písmen (znaků) je v angličtině?
- V češtině 40-42: 36 znaků (piš, jak slyšíš)
- V angličtině 40-44: 26 znaků
- Uvedená čísla fakticky předurčují povahu daného pravopisného systému: fonetický (čeština) vs. historický (angličtina: v zásadě má podobu pocházející ze 14. stol.)
- Spřežkové a diakritické systémy (nejen čeština)

Funkce pravopisných systémů

- Zaznamenávací – aby se to dobře psalo
- Vybavovací – aby se to dobře četlo
- Která funkce je důležitější? Vyváženost?
- Posílíme-li jednu funkci, oslabíme druhou a naopak
- Jaká je situace v praxi?
- Historické systémy (typicky angličtina, ze 14. stol.)
- Fonetické systémy (čeština, chorvatština, srbština)
- Spřežkové systémy – stará čeština, dnes zčásti i polština
- Lze získat značnou sumu za úspěšnou reformu anglického pravopisu (viz poslední vůle Bernarda Shawa, 1908, Simplified Spelling Society)
- Pravopisné systémy jsou velmi konzervativní (viz spory kolem reforem v němčině – řešil je něm. Ústavní soud), snaha o kompromisní řešení

Jazykové chyby

- Jazyková správnost a gramotnost – podmínka úspěšné komunikace a profesionální úspěšnosti obecně
- Pravopis není gramatika (popis struktury jazyka), je to soubor pravidel pro převod zvuků na znaky a tvorby srozumitelných textů
- Standard pro převod, ale patří sem také styl a interpunkce (rozhoduje o srozumitelnosti textu)
- Chyby: nedodržování standardů a norem
- Typy obvyklých pravopisných chyb v textech
- Překlepy (*prgram, studiijní*, …)
- Morfologické chyby – koncovky (*hloupejma, kerí*)
- Syntaktické chyby – shoda (psaní y/i), vazby (valence)

Chyby – pokr.

- Styl a stylistické chyby – např. *provedení nařízení,* opakování slov jako *prostě, teda, ten*
- Hlavním zdrojem styl. chyb je zpravidla chybná formulace myšlenek, snižuje srozumitelnost textu
- Typografické chyby – mezery, pomlčky, spojovníky, uvozovky, fonty, souhláskové předložky na konci řádků (např. *s, z, k, v, ,,,*)
- Dále tu je rozlišení: spisovnosti: nespisovnosti, formálnosti: neformálnosti

Opravování chyb v textech

- Snahou je, aby v našich textech bylo minimum chyb
- Lidé – korektoři nejsou dokonalí, v textech vždy zůstávají nějaké chyby i po opravách – korekturách
- Objevuje se proto snaha použít pro opravování chyb počítačů
- Vznikly pomůcky obsahující informace o jazykové správnosti – u nás Internetová jazyková příručka, v angličtině vybrané slovníky
- Další knižní příručky – slovníky – u nás norma: SSČ a SSJČ
- Elektronické nástroje odhalující překlepy a jiné chyby v textech: různé typy korektorů (gramatické, stylistické)
- Korektory dnes najdeme v produktech, jako jsou Microsoft Office (Word), Open Office aj.
- Co dovedou? Mohou být chytřejší než uživatel?
- Jen částečně – jsou rychlejší než lidé a konzistentnější

Pravidla českého pravopisu I

- Je český pravopis obtížný?
- Kombinace fonetického a historického principu, fonetický převažuje – piš, jak slyšíš, ang. pravop. je těžší k naučení
- Je dobré vědět, v čem se nejvíc chybuje?
- Stylistické chyby – cca 23 %, vznikají při snaze formulovat myšlenky, obecně jde o nesnadnou úlohu
- Interpunkční chyby – asi 20 %,
- Interpunkce zachycuje logické členění textu – srozumiteln.
- Ostatní chyby – překlepy, *y*/*i*, velká písmena, tvary *mne*/*mě*, koncovky, typografické chyby aj.
- Formulace standardů – reformy českého pravopisu – opakující se (spíše neúspěšné) pokusy

Příklad časté stylistické chyby

- „A tady jsme u toho. Čím méně se těch dat používá, protože se jim vyhýbáme, tak tím nepřispíváme k tomu, aby byla levnější. My potřebujeme změnit to klima v té společnosti také k tomu, aby ta data chtěli používat,“ doplnila ministryně (Nováková).
- Zdroj: https://www.idnes.cz/zpravy/domaci/novakova-data-mobil-internet-vtipy-wifi-benzin-nobelova-cena. A190219_143045
- Jde o chybné použití ukazovacích zájmen

Pravidla českého pravopisu II

- Pravidla č.pr. existují v knižní a elektronické podobě (poslední je sporné vydání ÚJČ – 1993), vykradené v.
- Internetová jazyková příručka (elektronická podoba pravidel č. p.) – je dílem ÚJČ a FI MU
- Má dvě části – slovníkovou a normativní
- Slovníková část pokrývá cca 110 000 českých slov
- IJP běží na serveru Centra ZPJ, denně přes 50 tis. přístupů (ukázka), též kniha Akademická přír. č. p.
- IJP obsahuje automatickou morfologii a je doplněna o dva normativní výkladové slovníky (SSČ a SSJČ)
- Používá se standardně ve školách a institucích v ČR

Korektory – pravopisné a jiné

- Jak jsou tyto nástroje konstruovány?
- Kolik slov má čeština? Kolik slovních tvarů?
- PSJČ – cca 250 000 základních tvarů slov (lemmat)
- Slovních tvarů v češtině je kolem 60 milionů
- Morfologická analýza a morfologický analyzátor jsou základem pro konstrukci českého korektoru překlepů
- Pro češtinu – morf. analyzátor Majka, cca 400 000 českých kmenů (vytvořen v Centru ZPJ, používá jej firma Seznam ve svém vyhledávači), v Praze – systém MORČE
- V české lokalizaci MS Wordu je použit korektor firmy Lingea (se sídlem v Brně)

Gramatické korektory I

- Dovedou opravovat slovní spojení v kontextu, např. *studentka šel do školi když pršeló.*
- Korektor překlepů ve Wordu značí chyby červeně, gramatický kor. je značí zeleně a opravy jen doporučuje
- Chyby v gramatické shodě a slovesných vazbách
- Upozorňují také na chyby v interpunkci (ne na všechny, úspěšnost je nejvýš do 60 %)
- Interpunkce odráží syntaktickou (logickou) strukturu věty – interpunkční korektor se vyvíjí v Centru NLP
- Gramatický korektor je k dispozici v české verzi Wordu (autoři: Oliva, Květoň, Petkevič)

Gramatické korektory II

- Též Grammaticon (od firmy Lingea), nízké pokrytí, kolem 40 %, falešné hlášky, neprodává se
- Co gramatické korektory nedovedou?
- Lze se na ně spolehnout jen v omezené míře
- Evaluační parametry: přesnost a pokrytí
- Přesnost (úspěšnost) se pohybuje do 70 %, ideálně zachycení všech chyb v textu (oblast UI)
- Principy fungování gramatických korektorů?
- Automatická syntaktická analýza – parsery (Synt)
- Heuristická pravidla – negativní příklady

Současný stav českého pravopisu I

- Poslední reforma čes. pravop. proběhla v r. 1993
- Pokus o tzv. „demokratizaci“ čes. pravopisu?
- Úprava psaní slov cizího původu (kurs/kurz)
- Reforma vedla k malé pravopisné válce (filos/zofie – láska k moudrosti vs. láska k temnotě)
- Výsledek: špinavý kompromis (větší fonetizace)
- nedomyšlenost reformy si lze ověřit – na velkých souborech textů – korpusech

Jaký je tedy současný stav? – je vidět, že norma je rozkolísaná – to je nejhorší možný výsledek

Současný stav čes. pravopisu II

- Korpusy (CzTenTen12, 5 miliard slovních tvarů) to potvrzují

  kurs: 65,360 vs. kurz: 581,913 výskytů, feminis/zmus: 11,895 vs. 332
- Rozkolísanost standardů je obecně nežádoucí, zhoršuje plynulost komunikace, možnost případné reformy?
- Bezbolestná úprava by např. byla sjednocení *ú*/*ů*
- Za problém se pokládají velká písmena, volnost je značná, lze pozorovat vliv angličtiny
- Pokud jde o *y*/*i*, situace pro úpravu není zralá
- Idiotské články na webu o zrušení ř – viz např. v červnu 2016:

  adresa: http://cool.iprima.cz/pravdivé-zprávy … !!!
- Případná inspirace slovenštinou, chytré řešení pro y/i
- Hledání konzistentních řešení – zatím není v dohledu

Předpokládaný vývoj českého pravopisu

- Institucí, která se stará o jazykovou kulturu v ČR, je Ústav pro jazyk český AV ČR (spolupráce na IJP)
- ÚJČ komunikuje s veřejností prostřednictvím Jazykové poradny a nyní též IJP
- Jejich aktuální přístup k problematice českého pravopisu je spíše liberální a zbytečně opatrný
- Pro studenty a absolventy FI je cílem, aby psali kultivovaně a bez chyb (bakalářské, diplomové práce)
- Kvalitní jazykový projev je jednou z cest k profesionálnímu úspěchu – nebezpečí fachidiotismu

Počítače a PJ I

- Komunikace mezi člověkem a počítačem je dnes primárně jednocestná (ale viz Pepper)
- Její kvalita ve skutečnosti závisí na tom, jak dobře uživatel zná programové vybavení svého počítače (jeho OS, pokus s OS Merlin, neujal se)
- Počítače s námi zatím nedovedou přirozeně komunikovat ve volné češtině (angličtině) – úloha je obtížná
- Potřeba dvoucestnosti – zpracování PJ je součástí umělé inteligence (viz obor UMI na FI)
- UI se týká modelování některých funkcí lidského mozku na počítači, nejen PJ, např. rozpoznávání obrazů, strojového učení, robotiky aj.

Počítače a PJ II

- UI a počítačové zpracování přirozeného jazyka
- Tři součásti – reprezentace znalostí o světě, inference (logika), gramatika (znalost PJ) – viz robot Pepper
- Dialogové systémy je musí obsahovat
- Turingův test, Eliza, chatboty, každoroční soutěž o Loebnerovu cenu (listopad 2018)
- Zpracování mluvené řeči – diktovací systémy
- Dovedou přepisovat zvuky na znaky – pro češtinu: Newton Technologies, Dictate 4…, lze je koupit za cca 7 tis. Kč (jde o aplikace ASR, TTS, např. je má Pepper)
- U těchto systémů ale ještě nejde o porozumění PJ

Jazykové inženýrství

- Elektronické slovníky české – nástroj DebDict
- Vícejazyčné elektronické slovníky – např. produkty firmy Lingea
- Google Translator (statistický přístup, neuronové sítě)
- České překladače: Eurotran, PC Translator – málo kvalitní, úspěšnost dosahuje asi 60-70 %
- Problematika strojového překladu obecně – těžká úloha
- Morfologické a syntaktické analyzátory
- Dialogové systémy, chatboty, dialog s Pepperem
- Porozumění přirozenému jazyku (extrakce informací v přirozeném jazyce – sumarizace, abstrakty)

Poznámka k předmětu VB000 (ZOS)

- Vztah k bc. a případně dipl. pracím – je to cvičení
- Na FI je celkem běžné, že bc. práce obsahují víc než 10 jazykových chyb, není to dobrá reklama pro stud. a FI
- Bc., dipl. A disertační práce jsou veřejně dostupné v ISu
- Práce s vyšším počtem chyb nebývají při obhajobě úspěšné, dostávají nižší hodnocení nebo je stud. neobhájí
- Ve VB000 se kolokvium získává za dvě písemné práce: Úvaha a Úvod k bc. práci – cvičení pro konkr. bc. práci
- Kvalita bc. prací výrazně závisí na vedoucích prací a oponentech, dobrý vedoucí bc. práce je klíčem k úspěšnému zvládnutí bc. či diplomové práce
- Vazba VB000 na předmět SBAPR – Bakalářská práce

Vazba VB000 na SBAPR II

- Garant SBAPR: doc. P. Matula
- Cíl – vytvoření konkrétní bc. práce
- Podmínka pro absolvování FI MU
- https://is.muni.cz/auth/predmety/predmet?vysl=2622030

Je třeba vybrat si téma bc. práce a jejího vedoucího (nabídka v ISu, vlastní témata)

Interaktivní osnova VB000

https://is.muni.cz/auth/el/1433/jaro2018/VB000/index.qwarp