Anotační manuál:

# Vytváření cílové hypotézy, tagování a vyhodnocování problematických jevů

V dokumentu jsou obsaženy jen jevy nezmíněné v anotačním schématu, případně jevy, které vyžadují podrobnější pojednání nebo specifikaci pro češtinu.

Materiál je zaměřen čistě na anotační rozhodnutí, nejsou zmíněny zásady týkající se práce s technikou (Falko, Exmaralda, MMAX2).

**Rozsah tokenů (TH1)**

Zkratky a emotikony považujeme za 1 token (např. P.S. = 1 token; :-) = 1 token).

Počet tokenů u kalendářních dat závisí na stavbě kalendářního údaje:

10. srpna = 1 token

10. srpna 2010 = 1 token

10. 8. 2010 = 1 token

10.08.10 = 1 token

desátého srpna 2010 = 3 tokeny

od 10. do 12. srpna = 4 tokeny (od /10./ do / 12. srpna)

10. srpna až 12. srpna = 3 tokeny

10. - 12. srpna 2010 = 1 token

**Interpunkční znaménka (TH1, EA1)**

Česká kodifikace ustavuje jako jediný možný zápis uvozovek formou kombinovaného znaku uvozovky dole + uvozovky nahoře (*„*xyz”). Při transkripci textů kandidátů jazykové zkoušky však nebylo rozlišováno, zda kandidát napsal první část znaku ve formě uvozovek dole. V korpusu se tedy vyskytuje pouze anglická varianta uvozovek (“xyz”) a není žádným způsobem zohledňována při chybové anotaci.

Čárky ve větě jednoduché nebo v souvětí doplňujeme dle platných pravidel českého pravopisu. Konstrukce tvořící samostatný větný člen neoddělený čárkou, např. *prázdninový kurz češtiny to je dobrý nápad*, však opravujeme následovně: *prázdninový kurz češtiny je dobrý nápad* a tagujeme jako nadbytečný subjekt (G\_Valency\_complnumb\_Ad).

**Diakritika (TH1, EA1)**

Tagy O\_Graph\_act (+\_O/\_Ad/\_Ch) užíváme v případě chybějící, přebytečné nebo zaměněné diakritiky u písmen: á, é, í, ý, ó, ú, ů, š, č, ž, ř, ě, ť, ď, ň

Např. e místo ě = O\_Graph\_act\_O; é místo e = O\_Graph\_act\_Ad; é místo ě = O\_Graph\_act\_Ch.

**Vokalizace předložek (TH1, EA1)**

Vokalizace předložek je považována za chybu ve výběru předložky. Př. *s sestrou → se sestrou*. Tag G\_Prep\_Ch.

**Velké písmeno u zájmen Ty/Tvůj, Vy/Váš (TH1, EA1)**

V české korespondenci se pro vyjádření úcty používá velké písmeno na začátku zájmen Ty/Tvůj, Vy/Váš. V případě, že kandidát převážně psal velké písmeno a ojediněle malé, opravujeme jednotně v celém textu na písmeno velké. Pokud kandidát užíval jednotně písmeno malé v celém textu, ponecháváme. Označujeme tagem pro nevhodně zvolené malé/velké písmeno: O\_Capit.

**Zvratná versus nezvratná osobní zájmena (TH1, EA1)**

Tag G\_Refl\_pronrefl\_Ch užíváme nejen pro označení záměny zájmen *se* a *si*, ale také zájmene *si/se* s odpovídajícím nezvratným zájmenem. Př. *těší mě na Tebe → těším se na Tebe.* V tomto problematičtějším, ale častém případě studentské chyby byla zvažována několikerá řešení, výsledné využívá jak tagu G\_Refl\_pronrefl\_Ch u zájmene, tak tagu G\_Agr pro chybnou osobu slovesa.

**Zvratná přivlastňovací zájmena (TH1, EA1)**

Užití nezvratných zájmen přivlastňovacích je dle gramatických pravidel češtiny nahrazováno zájmenem zvratným. Viz <http://prirucka.ujc.cas.cz/?id=630&dotaz=zvratn%C3%A9>: “Zvratné přivlastňovací zájmeno svůj užijeme tehdy, jestliže přivlastňovaná věc, event. osoba patří osobě/věci, která je ve větě agentem.” Př. *Mám hodně práce s mojí diplomkou.* Cílová hypotéza 1 bude: *Mám hodně práce se svojí diplomkou.* Tagujeme G\_Refl\_pronreflposs.

V případech nepřivlastňovacích zájmen tuto konkurenci neuvažujeme a ponecháváme studentovi nezvratné zájmeno. Neopravujeme tedy např. *Rád bych Tě pozval ke mně domů.* na *Rád bych Tě pozval k sobě domů.*

**Stupeň adjektiv a adverbií (EA1, EA2)**

Chyby týkající se stupně adjektiv a adverbií mohou být dvojího druhu:

a) gramaticky špatná forma: *jsem největší* (superlativ) *než ty* → *větší* (komparativ) *než ty*;

b) nevhodné užití: *bez svých oblíbenějších* (komparativ) *botiček* → *nejoblíbenějších* (superlativ)

Ad a) Gramaticky špatné formy jsou opravovány v rámci oprav na rovině gramatiky a ortografie a vyhodnocovány jako chyby ve flexi adjektiva, tag G\_Inflect\_adj.

Ad b) Gramaticky správné formy měníme až na TH2 a opatřujeme sémantickým tagem na EA2: V\_semdenot\_, V\_semimprec\_ či V\_wordform\_deriv.

**Kombinace tagů u syntagmat (TH1, EA1)**

V rámci syntagmat může docházet k různým kombinacím chyb v morfologických kategoriích substantiv, adjektiv a zájmen.

Uveďme na ukázku alespoň dva příklady kombinací a jejich tagování:

a) *vypravuj o prázdninové kurzy* *→ vypravuj o prázdninových kurzech*

b) *přijdu k vaše svatbu* *→* *přijdu na vaši svatbu*

V příkladě a) je správně volena předložka, shoda je správná mezi adjektivem a substantivem, nesprávné je však spojení předložky s daným pádem substantiva. Tagujeme všechny chybné složky fráze: G\_Morphol\_case\_wrong u adjektiva + G\_Morphol\_case\_wrong u substantiva.

V příkladě b) je chybná předložka, v cílové hypotéze je nahrazena předložkou *na*. Zájmeno není ve shodě se substantivem, pád substantiva byl vzhledem k původní předložce volen chybně. Značíme tagy: G\_Prep\_Ch + G\_Morphol\_case\_wrong u zájmene + G\_Morphol\_case\_wrong u substantiva.

**Infinitivy (TH1, EA1)**

Infinitiv může být užit nenáležitě ve třech různých případech:

a) *mám rád vařit*  *→ mám rád vaření* (G\_POS);

b) *já bych přemýšlet → já bych přemýšlel* (G\_Verb\_compl\_Ch);

c) *já mít hlad* *→ já mám hlad* (G\_Agr).

V případě a) vyhodnocujeme chybu jako záměnu slovních druhů, verba a verbálního substantiva. Toto řešení však nevolíme v případě spojení kolokace mít rád s infinitivem slovesa dokonavého, př. mám ráda podívat se (tento případ opraven dle kontextu na *ráda se podívám* - *mám* nadbytečné = G\_verb\_compl\_Ad; *podívat* = G\_Agr).

V případě b) se jedná o užití infinitivu místo činného příčestí, tj. komponentu složeného slovesného tvaru. Tagujeme v souladu s anotačním schématem jako G\_Verb\_compl\_Ch.

Případ c) vyhodnocujeme jako chybu shody, přesněji řečeno jako úplnou absenci morfologických kategorií osoba, způsob, číslo a čas.

**Složené versus jednoduché slovesné tvary (TH1, EA1)**

Nastat mohou 2 situace:

a) *budu jet → pojedu* (G\_Inflect\_verb + G\_Verb\_asp);

b) *budu psát → napíšu* (G\_Verb\_Asp).

V případě a) student užil v paradigmatu neexistující formu i špatný vid.

V b) byla studentova verze gramaticky v pořádku, avšak byl zvolen špatný vid.

Tyto případy nekombinujeme s tagem G\_Verb\_compl\_Ad.

**Chyby ve složených slovesných tvarech kondicionálu (TH1, EA1)**

Kondicionálový tvar pomocného slovesa *být* může student uvést chybně i v případě, že se jedná o spojení tohoto tvaru s výrazem spojkovým (*aby*). I v takových případech vyhodnocujeme chybu jako chybu shody. Př. *abychom mohl* *→ abych mohl,* tag G\_Agr.

**Vidové protějšky (TH1, EA1)**

Slovesa různých vidů nepovažujeme za různé lexémy.

Př. *zdravíš Martina ode mě* *→ pozdravíš → pozdravuj*

*Pozdravovat* je nedokonavý protějšek jak od *pozdravit*, tak *zdravit*. Student zde měl chybu ve vidu i způsobu. Po opravě vidu je výsledný tvar *pozdravíš* (G\_Verb\_asp), po opravě způsobu tvar *pozdravuj* (G\_Verb\_md).

**Slovosled a aktuální větné členění (TH1, EA1)**

Při opravování a anotování nesrovnalostí ve slovosledu bylo rozlišováno, zda se jedná o jevy gramaticky vázané, zpravidla v případě neplnovýznamových slov, nebo o jevy volné, zpravidla v případě plnovýznamových slov. Na základě tohoto rozlišení byl volen způsob anotace:

**A)** U slov příklonného charakteru (zvratná zájmena, krátké tvary osobních zájmen, pomocné sloveso být, -li, ...) v cílové hypotéze 1 (TH1) důsledně opravujeme slovosled, protože pozice je pevně zakořeněna a gramaticky dána (nejčastěji druhá pozice).

Tagujeme následujícími tagy: G\_Verb\_Compl\_Pos a G\_Refl\_pronrefl\_Pos.

Opravována je i špatná pozice předložek a spojek, značena je tagy: G\_Conj\_Pos, G\_Prep\_Pos.

Totéž platí v případě záporky “ne”, tag: G\_Neg\_neggen\_Pos.

A také u nesprávně umístěné interpunkce: O\_Punct\_Pos.

**B)** V případech slovosledu plnovýznamových a nepříklonných slov, který se nám zdá nepřirozený (například různé kombinace téma - réma, východisko - ohnisko), ale v zásadě možný, respektujeme pravidlo minimálních zásahů a ponecháváme.

**Spisovnost versus nespisovnost (TH1, EA1)**

Prvky typické pro nespisovné variety jazyka podléhají patřičným opravám již na TH1. Př. *s dětma*, cílová hypotéza *dětmi*, vyhodnoceno jako tvar neexistují v paradigmatu daného lexému (nelze říci přímo slovního druhu, neboť u některých substantiv je zakončení -ma v I pl. kodifikováno - rukama, nohama,...). Tagujeme G\_Inflect\_noun\_inexist.

**Cílová hypotéza, která vytváří chybu (TH1, EA1)**

Některé opravy studentova textu, tj. vytvoření cílové hypotézy, mají za důsledek nesoulad v rámci kontextu. Např. *pojedeme na Tatry*, oprava předložky vede k hypotetickému textu *pojedeme do Tatry*. Tento text je nutné z hlediska gramatického sladit, vytváříme tedy TH1 *pojedeme do Tater*, tagujeme však jen záměnu předložky (G\_Prep\_Ch), ne nesprávný pád. Nepřipisujeme studentovi chybu, kterou neudělal. Podobně v případech, kdy opravujeme interpunkci, netagujeme opravu následujícího velkého/malého písmene: *Ahoj! Jak se máš?* *→* TH1: *Ahoj, jak se máš?* EA1: pouze ortografický tag O\_Punct\_Ch.

**Algoritmus pro rozlišování ortografických, gramatických a některých lexikálních jevů (TH1, EA1, TH2, EA2)**

Teoretický koncept projektu Merlin dělí anotaci do dvou fází: 1) zpracování ortografických a gramatických chyb a jejich tagování (TH1 a EA1); 2) zpracování chyb vyšších jazykových úrovní (lexikální, sociolingvistická atd.; TH2 a EA2).

Některé chyby v češtině se pohybují na hranicích ortografie a gramatiky, jiné na hranici gramatiky a lexikologie apod. Odlišení mezi fázemi tedy není jednoduché. Pro mezianotátorskou jednotnost bylo nutné stanovit pomůcku, která by zajistila shodné vyhodnocování hraničních jevů, resp. fixovala intuitivní rozhodnutí jednotlivých anotátorů. K tomuto účelu byl vytvořen jednoduchý algoritmus (v barevných polích jsou uvedeny příklady):

*→ → → → → → → → → → → → → → → → → → → → → → → → → → → → → → → → →*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Jedná se o chybu v tvaroslovné charakteristice (TCH)?** | ANO | **Existuje v paradigmatu daného slova?** | ANO | všichni jsou spokojené | G\_Morphol\_gend\_wrong |  |  | chyba v morfologických kategoriích (u sloves G\_Agr) |
| NE | **Existuje v paradigmatech daného slovního druhu?** | ANO | večeřit | G\_Inflect\_verb\_inexist | chyba ve flexi |
| NE | kratke kalhoty | O\_Graph\_act\_O | některá z ortografických chyb |
| NE (kořeny, prefixy, části sufixů bez TCH) | **Slovotvorná motivace chyby?** | ANO | němečtina | V\_Wordform\_deriv |  |  | slovotvorná chyba |
| NE | tatinek | O\_Graph\_act\_O |  |  | některá z ortografických chyb |

*→ → → → → → → → → → → → → → → → → → → → → → → → → → → → → → → → →*

Algoritmus orientuje anotátora směrem k jazykové rovině, tj. ortografii, gramatice či slovotvorbě, a oblasti (viz poslední sloupec algoritmu), v níž je pak vybírán tag odpovídající charakteru chyby.

Př. *informace o vašim hotelu → vašem*:

Na otázky z algoritmu odpovídáme po řadě ANO - ANO. Půjde tedy o chybu gramatickou, a to chybu v morfologických kategoriích jmen (G\_Morphol\_wrong). Vzhledem k tomu, že tvar užitý kandidátem v paradigmatu daného slova existuje, doplňujeme tag o určení reflektující rozdíl mezi *vašim* a *vašem* v paradigmatu, tedy číslo a pád (\_case\_ + \_numb\_). Finální tagy jsou G\_Morphol\_case\_ wrong a G\_Morphol\_numb\_wrong.

Algoritmus je možné použít také opakovaně, tj. postupně “odbalovat” z tokenu chybné jevy.

Př. *ty mě navštivit u mě doma → navštívit → navštívíš*

Na otázky z algoritmu odpovídáme v případě -ti- po řadě ANO - NE - NE. Jedná se o ortografickou chybu v délce, tag O\_Graph\_act\_O. Výstup *navštívit* znovu zadáme do algoritmu se zaměřením na zakončení. Odpovědi jsou ANO - ANO, tedy u sloves chybná shoda, tag G-Agr.

Některé jazykovědné teorie považují i zakončení příslovcí za tvaroslovnou charakteristiku (neboť přiřazuje ke slovnímu druhu), v rámci anotace korpusu Merlin však v případě příslovcí o tvaroslovné charakteristice neuvažujeme.

Př. *cítí se moc hezkě* *→ cítí se moc hezky*

Na první otázku v případě příslovcí tedy odpovídáme NE. Ve druhé otázce volíme ANO, neboť zakončení -ě je v případě jiných příslovcí běžné. Student tedy udělal chybu slovotvorného rázu, tag V\_Wordform\_deriv.

Zároveň při užívání algoritmu platí jakési dílčí pravidlo upřednostňování jeho několikanásobného použití před jednorázovým, tj. upřednostňování vícevrstvých chyb.

Př. *hezke dopis → hezké dopis → hezký dopis*

Jedná se o chybu v tvaroslovné charakteristice, odpověď ANO. Krátké e není charakteristické pro paradigma daného slova ani slovního druhu, odpovědi NE, NE. Chyba je tedy ortografického rázu, ale nemůžeme rovnou zvolit záměnu grafému e za ý. Student totiž neprokázal ani deklinační dovednost. Volíme tedy grafém é a podobu *hezké* znovu podrobíme tázání; odpovědi ANO, ANO, ANO vedou k identifikaci morfologické chyby, tag G\_Morphol\_gend\_wrong (tvar *hezké* identifikujeme jako singulárový, tj. cílové hypotéze bližší, nevyžadující kupení více tagů G\_Morphol\_…).

Pravidlo upřednostňování několikanásobného užití algoritmu se ukázalo jako důležité v případech typu *s hezke knihou*, u nichž bychom v rámci ortografie museli zohledňovat chybu v záměně i absenci grafémů (e-ou).

**Slovní spojení, frazémy, idiomy (EA2)**

V případě pochybností, zda označit určitou sekvenci tagem (V\_FS\_colloc, FS\_idiom apod.) užívají anotátoři publikaci: Čermák, F. (ed.): Slovník české frazeologie a idiomatiky 1-4, případně některý z českých výkladových slovníků: Příruční slovník jazyka českého - PSJČ (<http://bara.ujc.cas.cz/psjc/search.php>), Slovník spisovného jazyka českého - SSJČ (<http://ssjc.ujc.cas.cz/>), Slovník současné češtiny - SSČ (<http://nechybujte.cz/>), případně Internetovou jazykovou příručku (<http://prirucka.ujc.cas.cz/>). Je také možné ověřit frekvenci kolokace v Českém národním korpusu - ČNK (http://korpus.cz/).

**Konektory (EA2)**

Pří zvažování tagů C\_Coh\_txtstruct a C\_Con\_accur využívají anotátoři listu konektorů (spojek a vztažných zájmen), který byl sestaven na základě českých gramatik (Mluvnice češtiny, Academia Praha 1987 - <http://stream.avcr.cz/ujc//mluvnice-cestiny-3.pdf?0.5591183473838413>; Štícha a kol.: Akademická gramatika spisovné češtiny, Academia, Praha 2013) a filtrace z Českého národního korpusu (SYN2010).

Vztažná zájmena (řazeno dle frekvence): *který, co, jenž, kdo, jaký, copak, čí, jakýpak, kdopak, kterýžto, an, cožpak, kterýž, kdos, kterýpak, čo, ký, jakýže, čípak, jakýž, kdožpak, jakýžto*

Spojky (řazeno dle frekvence): *a, že, ale, i, jako, když, aby, nebo, než, ani, však, protože, či, pokud, až, kdyby, takže, jestli, li, proto, ovšem, zda, zatímco, ať, jenže, neboť, vždyť, tak, jestliže, dokud, avšak, přestože, buď, anebo, jakmile, ačkoli, aniž, nicméně, nýbrž, tj, přičemž, byť, jednak, zato, , jelikož, tudíž, neboli, jenomže, ač, poněvadž, třebaže, coby, jak, , čili, kdežto, leč, jakožto, místo, aneb, jakož, nežli, zdali, jakoby, buďto, ledaže, plus, alias, co, pakliže, ni, namísto, ježto, a/nebo, div, seč, nechť, jakkoli, pročež, liž, sotvaže, jakože, pakli, zdaž, anžto, ačli, dokavaď, zdaliž, anobrž, pokaď, jakkoliv, buď-anebo + eventuálně, jedině, načež, nadto, naopak, natož, netoliko, respektive, toliko*

Vícečlenné konektory: *pokaždé když, do té doby, kdy, tehdy, když, od té doby, co, poté, co, potom, co, za těchto podmínek, za takových okolností, za těchto předpokladů, následkem toho, (a) v důsledku toho, v důsledku toho, že, navzdory tomu, přes to přese všechno, přes to všechno, a nakonec, a následkem toho, a pak , a potom, a proto, a přece, a přitom, a rovněž, a tak, a to, a vedle toho, a také, a tedy, a tudíž, a zatím, aby ne, a když, ale přesto, ale zato, i kdyby, ani kdyby, ani když, ať - ať, ať - anebo, ať - či, aťsi - aťsi, aťsi - nebo, až když, až na to, že, ba i, ba ani, ba dokonce, bez ohledu na to, zda/jestli/že, buď - anebo/nebo, buď - buď, byť i, co se týče, čím - tím, díky tomu, že, dílem - dílem, divže/div že, dotud - dokud, dřív než, hned - a zase, hned - hned, hned jak, i když, jak - tak, jak jen, jako (kdy)by, jako když, jak - tak, jednak - jednak, jedni - druzí, jen aby, ještě - (a) už, ještě ne - už, leda když, lépe řečeno, místo aby, napřed - potom, napřed - potom - konečně, následkem toho, následkem toho, ž, , natož aby, nejenže, nejen - ale (i), nejen(že) - nýbrž i, nejprve - potom - nakonec, neřkuli aby, než aby, pokud jde o, pokud ne, pokud se týče/týká, především - potom, předně - zadruhé, sice - ale/avšak, tady - tam - jinde, tak dlouho, dokud, tak - jak, tj., tu - tam, vzhledem k tomu, že, zde - tam (tu), na to, že*