Multilingual Platform for the European Reference Levels: Interlanguage Exploration in Context

Documentation of additional annotation decisions

Please cite as: MERLIN project, Documentation of additional annotation decision, 2014, http://merlin-platform.eu

Lifelong Learning Programme

This project has been funded with support from the European Commission. This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Multilingual Platform for the European Reference Levels: Interlanguage Exploration in Context

Documentation of additional annotation decisions – for Italian

Please cite as: MERLIN project, Annotation decisions, 2014, http://merlin-platform.eu

Teresa Knittel

Annotation manual: Annotation decisions for Italian

This paper documents the decisions taken during the annotation process in cases of problems and doubts for the Italian texts. Only decisions that are not included in the Annotation Scheme are documented here. Technical issues are not part of this document.

Explicit subjects (TH1, EA1)

As there are no clear rules for Italian, when the use of an explicit subject is an error (to be annotated) and when a stylistic matter (not to be annotated), we rely on the native-speaker competence for the decision when to correct and tag an explicit subject as valency error (G_Valency_complnumb_Ad). We try to intervene as little as possible, but correct sentences that are highly non-native like.

Missing agreement between implicit subject and verb (TH1, EA1, TH2, EA2)

If the missing agreement between an implicit subject and the verb is only inferable from the context and the sentence in isolation is grammatically correct, we do not correct the verb in TH1. We correct it in TH2 and tag it with C_Coh_ref in EA2 (i.e. no G_Agr or S_Form_addr).

> *E finalmente, quando [vorrei] Iniziare?* → TH1: *...vorrei iniziare?* → TH2: *...vuoi iniziare?* EA2: C_Coh_ref
>
> *Come stai? Che cosa fai? [Ha] molto lavoro?* → TH1: *...Ha molto lavoro?* → TH2: *...Hai molto lavoro?* EA2: C_Coh_ref

Agreement errors in past participle (EA1)

We use G_Agr for wrong agreement between subject and past participle, and G_Verb_compl_Ch for wrong agreement between object and past participle. In addition, we use the tag G_Clit_xxx if the predicate involves a wrong clitic.

> *Dove ci sono [scritto] le cose che...* → TH1: *Dove ci sono scritte le cose che...* EA1: G_Agr
>
> *Noi [le abbiamo aiutato]* → TH1: *Noi l'abbiamo aiutata* EA1: G_Verb_compl_Ch

Ambiguous learner errors (EA1)

In some cases the learner error is ambiguous, i.e. there are two or more possible interpretations of the error and therefore two or more possible tags could be applied. Usually we decide case by case. Here the more general decisions are documented.

Orthographic error vs. lexical error

Minor orthographic errors leading to the use of non-existing words or to a semantic error are tagged as orthographic errors and not as lexical ones (e.g. V_form_word_nonexist or V_semdenot_word).

**mo* → TH1: *ma*	EA1: O_Graph_graphgen_Ch
nove → TH1: *nuove*	EA1: O_Graph_graphgen_O
invio → TH1: *invito*	EA1: O_Graph_graphgen_O
e → TH1: *è*	EA1: O_Graph_act_O
a → TH1: *ha*	EA1: O_Graph_graphgen_O

Orthographic error vs. grammatical error (function words)

If the learner uses a wrong function word, we opt for grammatical error, even if only one letter is wrong compared to the right function word.

la → TH1: *le*	EA1: G_Morphol_numb_wrong (article)
al → TH1: *dal*	EA1: G_Prep_Ch
da → TH1: *di*	EA1: G_Prep_Ch

Orthographic error vs. grammatical error (G_Inflect_xxx_inexist) vs. lexical error (V_form_word_nonexist)

We opt for G_Inflect_xxx_inexist, if the form used by the learner does not exist and the error is found in the inflectional ending.

If the inflectional ending is partly or completely missing, we also use G_Inflect_xxx_inexist.

*le [*nazionalitè]* → TH1: *le nazionalità*	EA1: G_Inflect_noun_inexist
*la [*nazion]* → TH1: *la nazione*	EA1: G_Inflect_noun_inexist

If the error is in the word stem, we opt for orthographic error (see also above).

**mageremo* →TH1: *mangeremo*	EA1: O_Graph_graphgen_O
*la [*fabrica]* → TH1: *la fabbrica*	EA1: O_Graph_graphgen_O

Grammatical errors in formulaic sequences (TH1, EA1, TH2, EA2)

If there is at least one possible context in which the sentence is grammatically correct, we keep the learner version in TH1 and change it in TH2 to the version which is correct in the context. We tag it as form error in a formulaic sequence in EA2.

Presto! (at the end of a letter) → TH1: *Presto!* → TH2: *A presto!* EA2: V_FS_Form_O

(Possible context: *Quando vieni? – Presto!*)

If there is no context in which the sentence is grammatically correct, we correct it in TH1 and tag it in EA1 with the appropriate grammatical error tags. In addition, we use the appropriate tag for form errors in formulaic sequences in EA2.

giocare [al] calcio → TH1: *giocare a calcio* → TH2: *giocare a calcio* EA1: G_Art_Ad + EA2: V_FS_Form_Ad

Mi piacerebbe frequentare [sul] questo corso. → TH1: *...frequentare questo corso.* → TH2: *...frequentare questo corso.* EA1: G_Prep_Ad + G_Art_Ad, EA2: V_FS_Form_Ad

Wrong use of articles and partitive articles which can only be inferred from the context (TH1, EA1, TH2, EA2)

In cases of sentences in which the article or partitive article used is formally correct (i.e. the sentence is grammatically correct), but is not right in the global context, we correct the article or partitive article in TH2. No grammatical error tag can be applied in EA2, so the correction remains untagged.

mangiare i panini → TH2: *mangiare dei panini* EA2: no tag

frequentare i corsi di italiano → TH2: *frequentare dei corsi di italiano* EA2: no tag

TH1 corrections that become incorrect in TH2 (TH1, EA1, TH2, EA2)

If a wrong verb in TH1 (that can be corrected only in TH2) requires another preposition as the one written by the learner, we correct the preposition in TH1 and tag it in EA1. In the second step (TH2) we correct the verb and add the preposition required by the new verb, but in EA2 we tag only the semantic error of the verb.

partire [a] Milano → TH1: *partire per Milano* → TH2: *{andare a} Milano*
EA1: G_Prep_Ch; EA2: V_semdenot_word_1

Articulated prepositions (EA1)

We consider the componential character of articulated prepositions (preposition + article) and tag the element(s) of the articulated preposition which is (are) wrong (i.e. wrong preposition, wrong article).

giocare [al] calcio → TH1: *giocare a calcio* EA1: G_Art_Ad (not G_Prep_Ch)

frequentare [sul] questo corso. → TH1: *frequentare questo corso.* EA1: G_Prep_Ad + G_Art_Ad

Wrong article contraction/Non-existing articulated prepositions (EA1)

Non-existing contractions of preposition and article are tagged with G_Inflect_noun_inexist.

*[*nelli] programmi* → TH1: *nei programmi* EA1: G_Inflect_noun_inexist

Ambiguous contractions like *"nell Corriere"* are tagged as orthographic errors.

[nell] Corriere → TH1: *nel Corriere* EA1: O_Graph_graphgen_Ad

Wrong/non-existing gender of nouns (EA1)

If the learner uses a wrong, non-existing gender of a noun, we tag it with G_Inflect_noun_inexist, not with G_Morphol_gend_wrong. (G_Morphol_xxx_wrong is only used for forms that exist in the inflectional paradigm.)

We don't use V_form_word_nonexist, as the main problem is not that the word does not exist, but that the learner has chosen a wrong gender. So we tag a grammatical error.

**un'aiuta* → TH1: *un aiuto* EA1: G_Inflect_noun_inexist

**gruppa* → TH1: *gruppo* EA1: G_Inflect_noun_inexist

If the gender error of the noun leads to the use of an existing form of a different part of speech (e.g. verb form, adjective), but it is clear from syntax that the learner used the word as noun, we use G_Inflect_noun_inexist and not G_POS.

ho molto [lavora] → TH1: *ho molto lavoro* EA1: G_Inflect_noun_inexist

fare [passeggiati] → TH1: *fare passeggiate* EA1: G_Inflect_noun_inexist

Concordance (EA1)

If the learner uses the wrong gender for a noun and uses the same gender for article, prepositions ecc., only the gender error of the noun is tagged, because the learner is coherent in the choice of the gender.

un'aiuta → TH1: *un aiuto* EA1: G_Inflect_noun_inexist for noun, no error tag for article

Wrong use of interrogative and demonstrative adjectives (TH1, EA1, TH2, EA2)

If interrogative adjectives are used instead of demonstrative adjectives and viceverse, we tag it as semantic error in EA2, as they belong to the same part of speech and thus we can't apply G_POS. There is also no other appropriate grammatical tag. The error is therefore corrected in TH2 and not in TH1.

Puoi mi dire dove lavori adesso in quale ufficio e in [quella] città. → TH1: *...e in quella città?* → TH2: *...e in quale città?* EA2: V_semdenot_word_1

Vorrei sapere [quelle] parti ti piacciono. → TH1: *...quelle parti...* → TH2: *...quali parti...* EA2: V_semdenot_word_1

Puoi dirmi a [questa] ditta hai lavorato in tempo scorso. → TH1: *...in questa ditta...* → TH2: *...in quale ditta...* EA2: V_semdenot_word_1

Clitics (EA1, EA2)

If a learner omits a clitic and therefore an obligatory object, we tag it with both tags: G_Clit_O (clitic is missing) and G_Valency_complnumb_O (obligatory object is missing).

Every time that a clitic is missing or is wrong, it also involves a reference problem. Therefore we additionally use C_Coh_ref in EA2.

Ho dimenticato di [scrivere] prima. → TH1: *...di scriverlo prima.*
EA1: G_Clit_O + G_Valency_complnumb_O; EA2: C_Coh_ref

If the learner uses a wrong clitic, we use G_Clit_Ch and not G_Morphol_xxx_wrong (the latter tag includes pronouns and clitics are pronouns), because G_Clit_Ch is the more specific tag. Like above, we additionally use C_Coh_ref.

puoi [chiamarle] → TH1: *puoi chiamarla* EA1: G_Clit_Ch; EA2: C_Coh_ref

Errors in double pronouns (EA1)

We use G_Clit_Ch for tagging errors in double pronouns.

[*vi] la possiamo regalare* → TH1: *ve la possiamo regalare* EA1: G_Clit_Ch

Negation (TH1, EA1)

Missing negation words like in the following example are added in TH1 and tagged as G_Neg_neggen_O.

non fare [qualche cosa] di... → TH1: *non fare niente di...* EA1: G_Neg_neggen_O
(tagspan: [qualche cosa])

Apostrophes and Word boundaries (EA1)

If an apostrophe is missing and both words are attached to one another, we use O_Apostr_O and O_Wordbd_Merge.

**daccordo* → TH1: *d'accordo* EA1: O_Apostr_O + O_Wordbd_Merge

**anchio* → TH1: *anch'io* EA1: O_Apostr_O + O_Wordbd_Merge

Word boundaries: General tag O_Wordbd_Merge/Split and more specific tags G_Prep_Merge/Split and G_Art_Merge/Split (EA1)

O_Wordbd_Merge/Split is the general, overarching tag for all merges/splits, whereas G_Prep_Merge/Split and G_Art_Merge/Split are special cases of word boundary errors. That means that every G_Prep_Merge/Split and G_Art_Merge/Split is at the same time an O_Wordbd_Merge/Split. Therefore every word boundary error gets the tag O_Wordbd_Merge/Split and, if it involves prepositions or articles, G_Prep_Merge/Split or G_Art_Merge/Split has to be applied additionally.

In cases in which articles and prepositions are merged/splitted, we use G_Art_Merge/Split, not G_Prep_Merge/Split.

[perla] casa → TH1: *per la casa* EA1: O_Wordbd_Merge + G_Art_Merge

rispondi mi! → TH1: *rispondimi!* EA1: O_Wordbd_Split

Diacritics (TH1, EA1)

If an accent is used instead of an apostrophe, we use two tags: O_Apostr_O and O_Graph_act_Ad.

un [pò] → TH1: *un po'* EA1: O_Apostr_O + O_Graph_act_Ad

Punctuation (TH1, EA1)

We don't regard it as obligatory to use a comma between the greeting and the name of the author at the end of a letter / e-mail. So we don't add a comma in TH1 if the learner has not used one. Commas used by the learner are ok.

A presto Maria → TH1: *A presto Maria* EA1: no tag

A presto, Maria → TH1: *A presto, Maria* EA1: no tag

E-mail vs. *email*: We accept both forms (reference: Zingarelli 2013) and don't correct in TH1.

***Troncamento* and *elisione* (EA1)**

We tag cases of wrongly used *troncamento* or *elisione* as orthographic errors, as we don't have an appropriate grammatical tag for this (e.g. *troncamento* instead of *elisione*, *elisione* instead of *troncamento,* unnecessary *troncamento*).

[quel] esame → TH1: q*uell'esame* EA1: O_Graph_graphgen_O + O_Apostr_O

[bel'] lavoro → TH1: *bel lavoro* EA1: O_Apostr_Ad

[qual] bisogno avete? → *TH1: quale bisogno avete?* EA1: O_Graph_graphgen_O

Facultative elision (TH1, EA1)

We correct and tag if the learner's version sounds very non-native-like. If the facultative elision used by the learner is acceptable we don't correct and don't tag.

d'italiano → TH1: *d'italiano* EA1: no tag

di italiano → TH1: *di italiano* EA1: no tag

una impresa → TH1: *una impresa* EA1: no tag

un'impresa → TH1: *un'impresa* EA1: no tag

Errors in obligatory elisions are tagged as orthographic errors (see above).

d eufonica **(TH1, EA1)**

We only correct and tag if the learner's version sounds very cacophonous and thus non-native-like. In the other cases we leave the learner's version and don't tag. Tag to be used if correction is necessary: O_Graph_graphgen_O/Ad.

ed a fare → TH1: *ed a fare* EA1: no tag

Capitalization of *signora/-e/-i* **(TH1, EA1)**

We accept lower and upper case, but the learner should use it coherently in the text. In case the learner uses both variants in one text, we correct to the form mostly used in the text and tag the corrected forms with O_Capit.

Gentili signori → TH1: *Gentili signori* EA1: no tag

Gentili Signori → TH1: *Gentili Signori* EA1: no tag

Capitalization of personal pronouns and adjectives **(TH1, EA1)**

We accept lower and upper case, but the learner should use it coherently in the text. In case the learner uses both variants in one text, we correct to the form mostly used in the text and tag the corrected forms with O_Capit.

lei & *Lei, suo* & *Suo, voi* & *Voi, vostro* & *Vostro ecc.*

Capitalization of names of months **(TH1, EA1)**

We only accept lower case. We correct and tag both if the month is mentioned in the text and if it's in the beginning of the text, i.e. in the header.

5 [Agosto] 2010 → TH1: *5 agosto 2010* EA1: O_Capit

Errors in proper names **(EA1)**

We annotate errors in proper names only for:

a) names of countries, continents, oceans
b) names of Italian regions / federal states (endonyms)

Errors in other types of proper names (e.g. streets, companies, organizations) are corrected in TH1 but not tagged as error in EA1.

Nuova [Zealanda] → TH1: *Nuova Zelanda* EA1: O_Graph_graphgen_Ad

Via [Cavur] → *TH1: Via Cavour* EA1: no tag

Multilingual Platform for the European Reference Levels:
Interlanguage Exploration in Context

Documentation of additional annotation decisions – for Czech

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, ať si - ať si, ať si - 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*