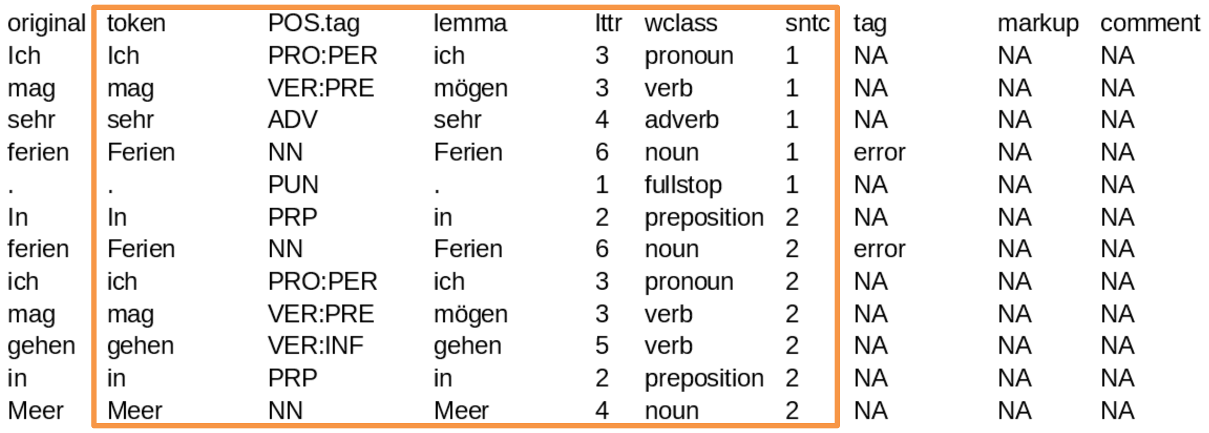
SWIKO: Error correction during transcription

Technical remarks

This document presupposes familiarity with the transcription program XMLmind and its tagging procedures. Refer to the document called “SWIKO_Instructions_XMLmind” for an introduction.

Introduction

During the transcription in XMLmind, the texts are prepared for automatic POS tagging:



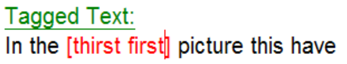
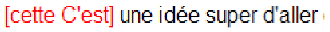
original	token	POS.tag	lemma	lttr	wclass	sntc	tag	markup	comment
Ich	Ich	PRO:PER	ich	3	pronoun	1	NA	NA	NA
mag	mag	VER:PRE	mögen	3	verb	1	NA	NA	NA
sehr	sehr	ADV	sehr	4	adverb	1	NA	NA	NA
ferien	Ferien	NN	Ferien	6	noun	1	error	NA	NA
.	.	PUN	.	1	fullstop	1	NA	NA	NA
In	In	PRP	in	2	preposition	2	NA	NA	NA
ferien	Ferien	NN	Ferien	6	noun	2	error	NA	NA
ich	ich	PRO:PER	ich	3	pronoun	2	NA	NA	NA
mag	mag	VER:PRE	mögen	3	verb	2	NA	NA	NA
gehen	gehen	VER:INF	gehen	5	verb	2	NA	NA	NA
in	in	PRP	in	2	preposition	2	NA	NA	NA
Meer	Meer	NN	Meer	4	noun	2	NA	NA	NA

During automatic POS tagging, the machine analyses the text and decides what word class a single word belongs to (e.g. a noun, verb, pronoun etc.) and, as a result, chooses the lemma (=unconjugated base form) of the word. This only works when the POS tagger recognises a word, i.e. when it is spelled correctly. The POS tagger we use (TreeTagger) also evaluates the surrounding words and makes assumptions such as “in English, a word with a capital letter is a name if it does not come after a period (.)” or “in French, a content word after a noun that is not a verb is an adjective”. However, these assumptions are not particularly reliable, especially with learner texts.

Word flexion, such as verb conjugation or noun declination, is not of interest to the POS tagger. Thus, “ich magst” or “tu jouons” will be correctly recognised as PRO:PER (pronoun: personal) and VER:PRE (verb: present). The same is true for nouns: “die Hund” or “les arbre” are correctly tagged as DET (article) and NN (noun).

Rules

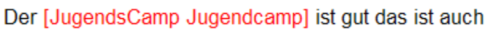
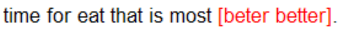
During the transcriptions, the learner texts are corrected so that the POS tagger can make mostly correct assumptions. The most important objective of this correction is that the words will be recognised as the correct word class. Thus, the correction mostly targets words that are misspelled, but only so far as the tagger would not recognise them correctly as being a word of the word class intended by the author. As mentioned before, a plural ending of an otherwise correctly spelled singular noun or an erroneous verb form do not have to be corrected. In contrast, a word that actually exists but is obviously a misspelled form of another word has to be corrected:

 Tagged Text: In the [thirst first] picture this have	While "thirst" (Durst/soif) does exist, it is clear from the overall text that the author had intended to write "first". Since the words are not of the same word class, the error has to be corrected.
 [cette C'est] une idée super d'aller	The word „cette" exists, but the author clearly intended the meaning "c'est" where the "t" is pronounced due to the following vowel.

In general: It is better to correct more often than not enough. All texts will later be tagged again in EXMARaLDA. So, if the POS tagger did not recognise a word or a tag is mistaken, this will later be corrected. But: the better the error correction is during transcription, the less work will have to be done later.

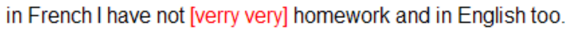
Obvious spelling errors

Words that are clearly spelled incorrectly will simply be corrected (*kanst* → *kannst*; *allways* → *always*).

 Der [JugendsCamp Jugendcamp] ist gut das ist auch	The spelling of the words is corrected, but not the fact that there are grammatical issues (*der Camp; *most better).
 time for eat that is most [beter better].	

In general, a correction should lead to a word of the correct word class, with as little changes as necessary!

- The target is always the word class probably intended by the author (*I life in …*→ *I live in …*). Any other mistakes should still be reflected if possible (*er sclaft* → *er schlaft*, not (as would be correct) *er schläft*).
- If there are several possibilities, the option that best reflects the mistake of the author should be chosen (*ich mochteste* → *ich mochte*, not *ich mochtest*, which contains an erroneous person ending even though it is closer to the original).
- If there are several distinct possibilities, the option that best reflects the context should be chosen (*wom* → *von*, instead of *wo* when it makes more sense in the text).
- Verbs are tagged according to their grammatical form (tempus, modus). Thus, if the author clearly intended a certain mode or time, correct verbs accordingly (*he leaved* → *he left*, even though *he leave* is less change).
- Words that mean something different but are of the same word class as the probable target word do not have to be corrected (*bet*, when the author meant *bed*; *chant* when the author meant *chat*).
- Misspelled words that mean something different but are of the same word class as the probable target word should only be corrected so they are spelled correctly:

 in French I have not [verry very] homework and in English too.	The spelling of the word "very" was corrected but not the fact that it

Capitalisation

Capital letters are recognised by TreeTagger before other criteria are taken into account. Thus, it is important to correct capital and lowercase letters if necessary. In general, the first word of the text and of each sentence (behind a period) must be capitalised. For this reason, if a sentence is clearly finished and a new one starts, either the student or the transcriber have to put a period (transcriber: with the tag "period"). The word that comes after the period has to be capitalized (if necessary with the tag "error").

If a word is written in capital letters only, it is transcribed as a "normal" word and tagged as "emphasis".

German

In German, any word with a capital letter is recognised as either a noun, a name or anything else if it comes directly behind a period (.). Thus, any word, that is not a noun or a name is only allowed a capital letter if it comes behind a period. Vice-versa, words behind a period that are not capitalized are not recognised correctly.

Rules:

- Capitalise nouns and names if the students did not.
- Capitalise the first word of each sentence (behind a period).
- Capitalise verbs or other words clearly used as nouns (Nominalisierungen), even if the nominalisation is not correct (*das Gesund* → *das Gesund* instead of **das Gesundheit*).
- Lowercase any other word except if it is the word behind a period (the first word of a sentence).

English

In English, any word with a capital letter is recognised as a name if it does not come directly behind a period. Vice-versa, words behind a period that are not capitalized are not recognised correctly. The first person pronoun *I* is only recognised when it is a capital i.

Rules:

- Capitalise names, nationalities and other word groups that are written with a capital letter (*spanish* → *Spanish*; *swisscom* → *Swisscom*; *thursday* → *Thursday*).
- Capitalise the first word of each sentence (behind a period).
- Capitalise the first person pronoun *I*.
- Lowercase any other word.

French

In French, any word with a capital letter is recognised as a name if it does not come directly behind a period. Vice-versa, words behind a period that are not capitalized are not recognised correctly.

Rules:

- Capitalise names (*swisscom* → *Swisscom*).
- Capitalise the first word of each sentence (behind a period).
- Lowercase any other word.

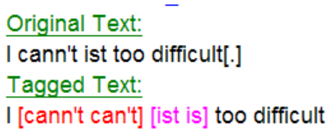
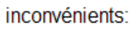
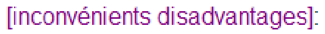
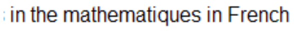
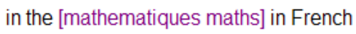
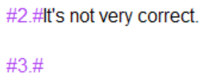
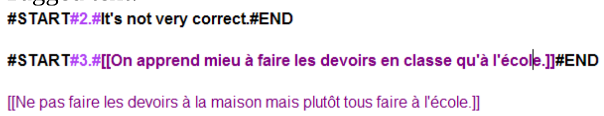
Words in another language than the text

Introduction

There are two separate groups of tags for words in another language than the overall text:

- foreign_Eng, foreign_Deu, foreign_Fra, foreign_oth: These tags are used when the students wrote words in a language other than the language of schooling (Eng: English, Deu: German, Fra: French). Usually, students tend to mix their two foreign languages, thus a French-speaking student writing in English might use German words in his/her text or a German-speaking student writing in French might use English words. The “foreign_oth” tag is used if there is good reason to believe that the student wrote a word in another than the three languages mentioned above. If possible, a comment can be added (in the comment tag) to specify the language used and a translation is given. If the word is unknown, it is not tagged!
- schlg_word and schlg_sentc: These tags are used when students write words or entire sentences in their language of schooling (e.g. French for students in the French-speaking part of Switzerland).

Words in other languages than the overall language of the text are not corrected but translated into the language of the text. Spelling (or other) mistakes in the original word are not corrected! If the tag “schlg_sentc” is used, no translation is given. The passage in the tag is treated like a comment (does not appear in word count, is not POS tagged etc.).

 Original Text: I cann't ist too difficult[.] Tagged Text: I [cann't can't] [ist is] too difficult	The German word “ist” in an English “text” was translated to “is” in the tagged text.
Original text:  inconvénients: Tagged text:  [inconvénients disadvantages]:	The French word “inconvénients” in the overall English text was translated to “disadvantages” in the tagged text.
Original text:  in the mathematiques in French Tagged text:  in the [mathematiques maths] in French	The French word “mathématiques” in the overall English text was translated to “maths” in the tagged text. The spelling error in “mathematiques” (no accent) was NOT corrected!
Original text:  #2.#It's not very correct. #3.# Tagged text:  #START#2.#It's not very correct.#END #START#3.#[[On apprend mieu à faire les devoirs en classe qu'à l'école.]]#END [[Ne pas faire les devoirs à la maison mais plutôt tous faire à l'école.]]	The purple passages in [[…]] are entire sentences written in French (the language of schooling) in an otherwise English text. They are not translated and they do not appear in the original text. Any spelling errors are kept as the author had written them (“mieu” instead of “mieux”).

Some students mark words in other languages in some way to show that they are aware that the word does not fit in the text. These markings are not transcribed.

Distinguishing between words in different languages

In some cases, it is difficult to decide which language tag is appropriate. In general, since the transcription is only the first step, it is most important that the word encountered by the TreeTagger be in the target language. Therefore:

- If students use a word that exists in the target language but obviously does not have the intended meaning because it is a cognate from another language, it is not tagged as belonging to that other language. It is, if necessary, corrected as usual (*Diese subjeckt gefällt mir* → *Diese Subjekt gefällt mir* in the sense of "school subject").
- If several options are possible, use the one that is as close as possible to the intended meaning.
- If several options are possible and the meaning is the same, use the following order of preference: Target language > Language of schooling > other foreign language in school > other languages, i.e. if a word might be in the target language or in the language of schooling, the target language is retained.
- When in doubt, use a dictionary to verify whether a word exists in a given language (German: duden.de, French: larousse.fr, English: www.merriam-webster.com).