VALICO-UD

Error Annotation Guidelines

Version 1.2

Elisa Di Nuovo (elisa.dinuovo@unito.it), 2021

Introduction

In designing this error coding system, we tried to follow as much as possible the requirements stated by Granger (2003). Using her words, an error coding system to be fully effective should be: 1. informative but manageable; 2. reusable; 3. flexible; 4. consistent (see her paper for a detailed description). Thanks to its structure, this error coding system easily copes with these requirements. In fact, it is flexible (it is potentially expandable), it is informative (the more letters the more fine-grained the annotation) but manageable (only three positions and no more than eighteen letters to remember), it is reusable (it can be applied to languages different than English or Italian), and consistent (the same phenomenon cannot be tagged with different codes).

Our error coding system is based on Nicholls (2003). Its tags consist of maximum three letters. Following the principle explained by Nicholls in her article, "the first letter represents the *general type of error* (e.g., wrong form, omission), while the second letter identifies the *word class of the required word*". We added a third letter which represents the grammatical category involved (Simone 2008), to provide a finer-grained description of the error.

With this principle in mind, all errors are encoded using a fixed set of letters which can occur in the first, second and third position.

0. General instructions

The error-tagged sentence is added to the err field of both learner sentence (LS) and target-hypothesis (TH) CoNLL-U files, as shown in the example below.

# sent_id = 1
# text = Più avanti c'è una signora seduta un po' spaventata, magari anche un po' arrabbiata, perché un cane è passato di corsa mandando per aria il tavolo dove lei mangiava: per terra c'è la bottiglia e per aria i panini.
# err = <SIM><i>Piu</i><c>Più</c></SIM> avanti c'è una signora seduta un <SEM><i>po</i><c>po'</c></SEM> spaventata, magari anche un <SEM><i>po</i><c>po'</c></SEM> arrabbiata<SPM><c>,</c></SPM> <SIM><i>perche</i><c>perché</c></SIM> un cane è passato di corsa mandando per <RN><i>area</i><c>aria</c></RN> il tavolo dove <IAG><i>lui</i><c>lei</c></IAG> mangiava<SPR><i>,</i><c>:</c></SPR> per terra c'è la <SB><i>botiglia</i><c>bottiglia</c></SB> e per aria i panini.

Example 1: the fields contained in the TH CoNLL-U file.

The correspondent LS of Example 1 is represented in Example 2.

# sent_id = 1
# text = Piu avanti c'è una signora seduta un po spaventata, magari anche un po arrabbiata perche un cane è passato di corsa mandando per area il tavolo dove lui mangiava, per terra c'è la botiglia e per aria i panini.
# err = SIM><i>Piu</i><c>Più</c></SIM> avanti c'è una signora seduta un <SEM><i>po</i><c>po'</c></SEM> spaventata, magari anche un <SEM><i>po</i><c>po'</c></SEM> arrabbiata<SPM><c>,</c></SPM> <SIM><i>perche</i><c>perché</c></SIM> un cane è passato di corsa mandando per <RN><i>area</i><c>aria</c></RN> il tavolo dove <IAG><i>lui</i><c>lei</c></IAG> mangiava<SPR><i>,</i><c>:</c></SPR> per terra c'è la <SB><i>botiglia</i><c>bottiglia</c></SB> e per aria i panini.

Example 2: the fields contained in the LS CoNLL-U file.

Let us see how to perform the error annotation which must be included in the **err** field.

1. The error code and its positions

In this section and its subsections, we will see in detail what are the letters that can occur in the first, second and third position of the error code and explain their meaning.

1.1 First Position

The letter in first position represents the general type of error. We identified eight types: derivation, form, inflection, missing, replacing, spelling, unnecessary, word order.

D = Derivation	1. it signals all that errors involving a word which was wrongly derived (further specified with the letter in the second position indicating the word class): · <DJ><i>enamorati</i><c>innamorati</c></DJ> · **<DT>**<SB><i>deill</i><c>deil</c></SB><i>deil</i><c>del</c>**</DT>** (note that we consider a *preposizione articolata* as a preposition, but when gender or number are involved with a *preposizione articolata*, we then mark the error in the determiner → e.g., IDG del → della)
F = Form	1. It signals all that errors involving a word to be changed because of its wrong form (further specified with letters in second and in third position – the latter is optional): · <FA><i>urlare a me</i><c>urlarmi</c></FA> · <FJ><i>bel</i><c>bell'</c></FJ>; 2. it is used in combination with L in third position for indicating (non)adapted loan words: · <FNL><i>cofferi</i><c>valigie</c></FNL> 3. it is used in contraposition of I_G when there is a problem of form selection and not with inflectional phenomena (e.g., gender or number): · <FD><i>i</i><c>gli</c></FD> (both masculine and plural) · <IDG><i>le</i><c>i</c></IDG> (both plural, but different gender)

I = Inflection	1. It catches all the errors related to conjugation and declension (further specified with the letter in the second position indicating the word class and in the third position indicating the grammatical category involved): · <IAG><i>lui</i><c>lei</c></IAG> · <IV><i>interposato</i><c>interposto</c></IV> · <IX><i>eraveno</i><c>erano</c></IX> · <IDN><i>sulla</i><c>sulle</c></IDN> · <IJG><i>bella</i><c>bello</c></IJG> (cfr. <FJG><i>bel</i><c>bello</c></FJG>) 2. If the learner uses a gender of number morpheme with an invariable word, we mark this kind of overextension error with U in first position and the letter for the specific phenomenon in third position (indicating unnecessary, PoS, phenomenon): · <UJG><i>enormo</i><c>enorme</c></UJG> (see third position for other examples)
M = Missing	1. It signals errors involving a missing word. · <MAX><i>sono</i><c>ci sono</c></MAX> 2. When it is a missing involving spelling errors—such as diacritics—we use S_M (indicating spelling, phenomenon, missing): · <SEM><i>po</i><c>po'</c></SEM> · <SIM><i>citta</i><c>città</c></SIM>
R = Replace	1. It signals errors involving a word to be replaced (further specified with the letter in second and in third position – the latter is optional): · <RD><i>l'</i><c>un</c></RD> (definitness error) · <RAR><i>che</i><c>in cui</c></RAR> 2. Used with S in first position it indicates the replacement of other phenomena, such as diacritics: · <SIR><i>perchè</i><c>perché</c></SIR> (see third position for other examples)

S = Spelling	1. It signals errors involving spelling issues: · <SC><i>que</i><c>che</c></SC> · <SN><i>statione</i><c>stazione</c></SN> · <SNL><i>instinto</i><c>istinto</c></SNL> · <SNS><i>Barbone</i><c>barbone</c></SNS> · <SB><i>botiglia</i><c>bottiglia</c></SB> · <SV><i>ligitò</i><c>litigò</c></SV> (cfr. <IV>) · <SX><i>a</i><c>ha</c></SX> · <SIM><i>e</i><c>è</c></SIM> (no <SX>)
U = Unnecessary	1. It signals an unnecessary word or with S in first position, it indicates other phenomena, such as spelling errors (e.g., diacritics, apostrophes): · <SEU><i>un'</i><c>un</c></SEU> abbraccio · <SIU> <SB> <i> piùtostto </i> <c> piùttosto </c> </SB> <i> piùttosto </i> <c> </c> piuttosto </SIU> · per liberare <UT><i>a</i><c></c></UT> la ragazza See I in first position at point 2 for overextension errors.
W = Word order	1. It signals a word in a wrong position: · <WA><RA><i> guardava lui</i><c> guardava lo</c></RA><i> guardava lo</i><c>lo guardava</c></WA>

1.2 Second Position

The letters in the second position indicates the word class or also other phenomena if a specific letter is in first position (e.g., SB indicates spelling double consonants).

Following Nicholls (2003), A stands for pronoun, C for conjunction, D for determiner, J for adjective, N for noun, T for adposition, V for verb. We adhere to Universal Dependencies PoS annotation choices (https://universaldependencies.org/u/pos/). To these we added B for double consonants, E for apostrophes, I for diacritics (accent), O for interjections, P for punctuation, R for adverbs, X for auxiliaries and W for generic words.

B = douBle consonants	It is used in combination with S in first position: · <SB><i>andrano</i><c>andranno</c></SB>

E = apostrophE	It is used in combination with S in first position and M, R or U in third position: · <SEU><i>un'</i><c>un</c></SEU> eroe · <URG><SEM><i>daccorda</i><c> d'accorda</c></SEM><i>d’accorda</i><c>d’accordo</c></URG> · <SER><i>pò</i><c>po’</c></SER> · <SEM><i>un </i><c>un’</c></SEM>eroina · <SEM><SIM><i>Ce</i><c>Cè</c></SIM><i>Cè</i><c>C’è</c></SEM> (note that we do not treat a missing apostrophe as a tokenization problem)
I = dIacritic	It is used in combination with S in first position and M, R or U in third position: · <SIM><i>perche</i><c>perché</c></SIM> · <SIM><SIU><i>pèro</i><c>pero</c></SIU><i>pero</i><c>però</c></SIM> · <SIR><i>perchè</i><c>perché</c></SIR>
W = generic Word	It is used in combination with S in first position and T in third position, when it is not possible to signal only one word class: · <SWT><i>perterra</i><c>per terra</c></SWT> It is used in combination of W in third position when a single word is replaced with more than one word: · sei <RWW><i>sicura</i><c>al sicuro</c></RWW> adesso.

1.3 Third Position

The third letter encodes the information about the grammatical categories involved and other phenomena (not addressed in the first or second position).

A = aspect	It is used only with I in first position: · <IVA><i>ha letto</i><c>stava leggendo</c></IVA> (solo perifrasi progressiva)
B = co-occurring tense and mood	It is used only with I in first position: · <IVB><i>ha fatto</i><c>facesse</c></IVB> · <IVB><i>portando</i><c>portava</c></IVB> · <IVB><i>bisognava</i><c>avesse bisogno</c></IVB> · ha fatto <IVB><i>caduto</i><c>cadere</c></IVB>

G = gender	It is used only with I or F in first position: · <ING><i>aiuta</i><c>aiuto</c></ING> · <FDG><i>ai</i><c>agli</c></FDG> · <UJG><i>enormo</i><c>enorme</c></UJG> · <SEM><IRG><i>daccorda</i><c>daccordo</c></SRG><i>daccordo</i><c>d'accordo</c></SEM> · <IVG><i>successo</i><c>successa</c></IVG>
L = calque or linguistic transfer	It is used only with F, R, I or S in first position (also con I): · <SNL><i>instinto</i><c>istinto</c></SNL> · <FAL><i>que</i><c>che</c></FAL> · aiuta<RAL><i>me</i><c>mi</c></RAL> · <RNL><i>i suoi ombrelli</i><c>le sue spalle</c></RNL> · <FTL><i>de</i><c>di</c></FTL> · <IVL><i>sapebba</i><c>sapevo</c></IVL> · <FVL><i>golpeare</i><c>colpire</c></FVL> · <INL><i>drinks</i><c>drink</c></INL>
M = mood	It is used only with I in first position: · <IVM><i>aveva</i><c>avesse</c></IVM> · <IVB><i>mettere</i><c>messo</c></IVB>
N = number	It is used only with I in first position: · <IJN><IJG><i>altre</i><c>altra</c></IJN><i>altra</i><c>altro</c></IJG> · <IVN><i>occupato</i><c>occupati</c></IVN> · <IJN><i>stesso</i><c>stessi</c></IJN> · <IDN><i>sulla</i><c>sulle</c></IDN>
O = collocation	· cassetta degli <RNO><i>arnesi</i><c>attrezzi</c></RNO>
P = person	It is used only with I in first position: · <IVP><i>esci</i><c>esce</c></IVP>

O = gerundive form instead of relative clause	It is used only with I in first position: · il ladro <IVO><i>rubando</i><c>che sta rubando</c></IVO>
S = capitalization	It is used only with S in first position: · <SCS><i>ma</i><c>Ma</c></SCS> · <SAS><i>MI</i><c>Mi</c></SAS> · Il <SNS><i>Barbone</i><c>barbone</c></SNS>
T = Tense or Tokenization	When it is used with I in first position, it means **TENSE**: · <IVT><i>l'ha saputo</i><c>lo sapeva</c></IVT> (not IVA, although aspect is involved) Note that we treat as tense errors those involving a missing auxiliary, e.g., mangio → ho mangiato, marked IVT no MX; but we mark MX in cases like this: visto → ho visto When it is used with S in first position, it means **TOKENIZATION** and signals errors involving a word which is wrongly formed due to a missing/unnecessary space (such as missing and replacing, it is signalled in the third position because we consider these errors as spelling errors). It can be followed by W indicating a general word or by a letter indicating a specific word class): ● <SWT><i>perterra</i><c>per terra</c></SWT> ● <STT><i>di l'</i><c>dell'</c></STT> (as we said above, we consider a *preposizione articolata* as a preposition, except when Gender of Number are involved)
W = multi-Word expression	· <FWW><i>una borsa di patatini</i><c>un sacco di patate</c></FWW>
X = existential construction	· <SVScascade>[1]<MAX> <i> Sono </i><c>Ci Sono</c></MAX><i>Ci Sono </i> <c>Ci sono</c></SVScascade> · <MAX><i></i><c>ci</c></MAX> sono

[1] See CASCADE ERROR

1.4 Special Tags

- **RSE** → replace subordinate explicit/implicit
  Finse <RSE><MC><i>aveva paura</i><c>che aveva paura</c></MC><i>che aveva paura</i><c>di avere paura</c></RSE> di un <DN><i>rapito</i><c>rapimento</c></DN>.
  ma la ragazza molto <SB><SJ><i>arraviata</i><c>arrabiata</c></SJ><i></i>arrabiata</i><c>arrabbiata</c></SB> mi ha detto **<RSE><i>che</i><c>di</c></RSE>** non <WA><RV><i>mi dichiarare</i><c>mi immischiare</c></RV><i>mi immischiare</i><c>immischiarmi</c></WA>
  <MD><i>in</i><c>nei</c></MD> loro affari.

- **RS** → replace subordinate
  Quando la signorina ha visto Marco il quale non si muoveva più, non era contenta! → quando la signorina ha visto che Marco non si muoveva più
  <RS><i>Marco il quale non si muoveva più</i><c>che Marco non si muoveva più</c></RS>

- **RCS** → replace coordinate to subordinate
  Sono ritornata <MD><i>di</i><c>dalla</c></MD> vacanza <RCS><i>e</i><c>in cui</c></RCS> ho traversato tutta la Francia, cioè 95 kilometri con la macchina durante 9h30 (durante – per – in)

- **RRC** → replace adverbs correlatives
  tra più gente, più storie ridicole → quanta più gente, tante più storie ridicole
  <RRC><i>tra più gente, più storie ridicole</i><c>quanta più gente, tante più storie ridicole</c></RRC>

We also included a catch-all code (CC: complex error), such as in Nichols (2003), to cover multiple complex errors. We decided to use it only in exceptional cases. At the moment it has been used only once:

- che la donna <CC><i>rovesciare l'eccedenza equipaggia la spalla</i><c>veniva portata sulla spalla</c></CC>

For all the cases in which it is possible to use nested XML tags (see next section), we opt for this alternative.

Note that we defined specific letters for some recurrent co-occurrences (such as B in third position).

2. Nested tags

We use nested tags to show the necessary steps to go from the LS to the TH version, for example:

<IVT><SV><SB><i>speggia</i>**<c>spegia</c></SB><i>spegia</i><c>spiega</c></SV></i>spiega</i>**<c>ha spiegato</c>**</IVT>**

In nested tags, inside the <i></i> tag you have to write the word contained in the previous <c></c> tag (**see these colours in the example above**). In the example above note that in cases in which there are two different spelling errors (in the example, transposition and double consonants) we mark both the errors.[2]

Dealing with clitics, they will be corrected along with the verb, because in some cases orthographical changes are necessary. In the first version, we treated them differently according to this distinction.[3]

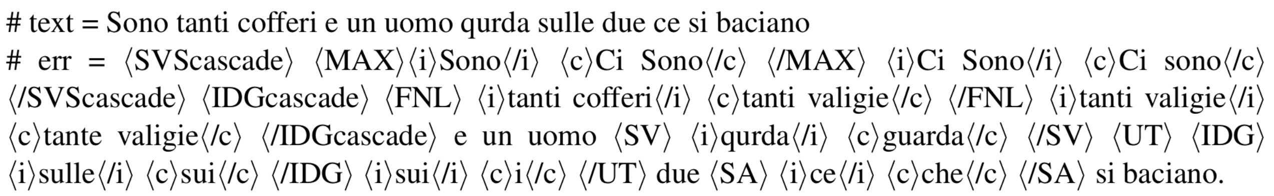
2.1 Cascade errors

If you are dealing with a CASCADE ERROR (e.g. you correct an error changing the gender of that word and in this way you cause an agreement error), you correct both the errors and signal that one is a cascade error adding "cascade" after the normal error tag, for example:

<IDGcascade><FNL><i>**tanti** cofferi</i><c>**tanti** valigie</c></FNL><i>**tanti** valigie</i><c>**tante** valigie</c>**<IDGcascade>**[4]

2.2 Hierarchical order

To ensure consistency across different annotators when dealing with nested tags, we provided a hierarchical order in the error annotation guidelines. Broadly speaking, we organized the errors in a pyramid, where in its base we have mechanical errors (i.e., tokenization, capitalization, spelling, and punctuation); in proceeding towards the apex, we find morphological (derivation and inflection), lexical and morpho-syntactic (form and replace), and syntactic (missing, unnecessary and word order) errors. For example, following this hierarchical order, mechanical errors should be corrected before a syntactic error. However, cascade errors make an exception and change the correction order, as we see in the example below in which we have a cascade capitalization error (SVS#) caused by a missing pronoun error (MAX) and a cascade inflection error (IDG#) due to a lexical error (FNL).



text = Sono tanti cofferi e un uomo qurda sulle due ce si baciano
err = ⟨SVScascade⟩ ⟨MAX⟩⟨i⟩Sono⟨/i⟩ ⟨c⟩Ci Sono⟨/c⟩ ⟨/MAX⟩ ⟨i⟩Ci Sono⟨/i⟩ ⟨c⟩Ci sono⟨/c⟩ ⟨/SVScascade⟩ ⟨IDGcascade⟩ ⟨FNL⟩ ⟨i⟩tanti cofferi⟨/i⟩ ⟨c⟩tanti valigie⟨/c⟩ ⟨/FNL⟩ ⟨i⟩tanti valigie⟨/i⟩ ⟨c⟩tante valigie⟨/c⟩ ⟨/IDGcascade⟩ e un uomo ⟨SV⟩ ⟨i⟩qurda⟨/i⟩ ⟨c⟩guarda⟨/c⟩ ⟨/SV⟩ ⟨UT⟩ ⟨IDG⟩ ⟨i⟩sulle⟨/i⟩ ⟨c⟩sui⟨/c⟩ ⟨/IDG⟩ ⟨i⟩sui⟨/i⟩ ⟨c⟩i⟨/c⟩ ⟨/UT⟩ due ⟨SA⟩ ⟨i⟩ce⟨/i⟩ ⟨c⟩che⟨/c⟩ ⟨/SA⟩ si baciano.

[2] Note that in the first version of the error tagset we annotated only the most specific error, in this case only SB and not SV → <SB><i>speggia</i><c>spiega</c></SB>.
[3] sgridar<RAL><i>me</i><c>mi</c></RAL> and <FVcascade><FA><i>**gridare** a me</i><c>**gridare**mi</c></FA><i>**gridare**mi</i><c>**gridar**mi</c></FVcascade>.
[4] In the first version, cascade errors were marked with a hashtag after the error code.