Error Annotation Guidelines

VALICO-UD

Version 1.2

Elisa Di Nuovo, 2019

**Error Annotation Guidelines**

**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 in order 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 cope 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 position 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). We structured it in order to have up to three letters composing each error code. 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), in order 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 learner sentence (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.

**1.1 First Position**

|  |  |
| --- | --- |
| **D = Derivation** | 1. 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>  · <DV><i>interposato</i><c>interposto</c></DV>  · <DX><i>eraveno</i><c>erano</c></DX>  · **<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 is involved with an articulated adposition, we then mark the error in the determiner 🡪 IDG del 🡪 della) |
| **F = Form** | 1. signals all that errors involving a word to be changed because of its wrong form (further specified with the letter in the 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>;   1. it is used in combination of some phenomena specified in third position:   · <FNL><i>cofferi</i><c>valigie</c></FNL>   1. Used in contraposition of I\_G when there is a problem with not with the gender:   · <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. 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>  · <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>)   1. If the learner uses a masculine morpheme with an invariable word, we mark this kind of overextension with U\_G:   · <UJG><i>enormo</i><c>enorme</c></UJG>  3. see third position for other examples; |
| **M = Missing** | 1. signals errors involving a missing word. 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>  · <SEM> <SIM> <i> Ce </i> <c> Cè </c> </SIM> <i> Cè </i> <c> C’è </c> </SEM>  · <MAX><i>sono</i><c>ci sono</c></MAX> |
| **R = Replace** | 1. signals errors involving a word (or other phenomena, such as diacritics) 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>  · <SIR><i>perchè</i><c>perché</c></SIR>  (see third position for other examples) |
| **S = Spelling** | 1. 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. <DV>)  · <SX><i>a</i><c>ha</c></SX>  · <SIM><i>e</i><c>è</c></SIM> (no <SX>) |
| **U = Unnecessary** | 1. signals an unnecessary word or other phenomena (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 2. for overextension errors. |
| **W = Word\_order** | 1. 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 indicate the word class and other phenomena.

|  |  |
| --- | --- |
| **A = Pronoun** | <https://universaldependencies.org/u/pos/PRON.html>  <https://universaldependencies.org/u/pos/NUM.html>  (if it is pronoun else see M below)  <https://universaldependencies.org/u/pos/PART.html> (negation particle –> pronoun) |
| **B = douBle consonants** | used in combination of S in first position:  · <SB><i>andrano</i><c>andranno</c></SB> |
| **C = Conjunction** | (no coordinating/subordinating conjunction distinction)  <https://universaldependencies.org/u/pos/CCONJ.html>  <https://universaldependencies.org/u/pos/SCONJ.html>  · <TC><FA><i>Per quello</i><c>Per ciò</c></FA><i>Per ciò</i> <c>Perciò</c></TC>  · <TC><MI><i>per che</i><c>per ché</c></MI><i>per ché</i><c>perché</c></TC>  · <MC><i></i><c>e</c></MC> |
| **D = Determiner** | <https://universaldependencies.org/u/pos/DET.html>  · <RD><i>la</i><c>una</c></RD>  · <WD><i>mia</i></WD> colpa <WD><c>mia</c></WD> |
| **E = apostrophE** | used in combination of U and M in first position:  · <UE><i>un'</i><c>un</c></UE> eroe  · <URG><SEM><i>daccorda</i><c> d'accorda</c> </SEM><i>d’accorda</i><c>d’accordo</c></URG>  (note that we do not treat a missing apostrophe as a tokenization problem) |
| **I = dIacritic** | used in combination of M, R and/or U in first 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> |
| **J = adJective** | <https://universaldependencies.org/u/pos/ADJ.html>  · <IJN><i>stesso</i><c>stessi</c></IJN>  · <SJ><i>lungha</i><c>lunga</c></SJ>  · <RJ><SJ><i>musculare</i><c>muscolare</c></SJ><i>muscolare</i><c>muscoloso</c></RJ> |
| **M = nuMeral** | <https://universaldependencies.org/u/pos/NUM.html>  (see A if it is a pronoun) |
| **N = Noun** | <https://universaldependencies.org/u/pos/NOUN.html>  · <FNL><i>prince</i><c>principe</c></FNL>  · <RNL><i>banca</i><c>panca</c></RNL>  · <SB><SNT><i>mmarcia piede</i><c>mmarciapiede</c></SNT> <i>mmarciapiede</i><c>marciapiede</c></SB> |
| **O = interjectiOn** | <https://universaldependencies.org/u/pos/INTJ.html> |
| **P = Punctuation** | <https://universaldependencies.org/u/pos/PUNCT.html>  · <SPM><i></i><c>,</c></SPM>  · <SPU><i>,</i><c>\_</c></SPU> |
| **R = adveRb** | <https://universaldependencies.org/u/pos/ADV.html>  · <SRT><i>in somma</i><c>insomma</c></SRT>  · <UR><i>rapidamente</i><c></c></UR> Giulio ha capito subito |
| **T = adposiTion** | <https://universaldependencies.org/u/pos/ADP.html>  · <RT><i>della</i><c>alla</c></RT>  · <FTL><i>En</i><c>In</c></FTL>  · <UT><i>di</i><c>\_</c></UT> |
| **V = Verb** | <https://universaldependencies.org/u/pos/VERB.html>  · <DV><i>svolgata</i><c>svolta</c></DV>  · c'<IVN><i>era</i><c>erano</c></IVN>  · <IVT><i>dice</i><c>ha detto</c></IVT> |
| **W = generic Word** | used in combination of T in first position when it is not possible to signal only one word class:  · <SWT><i>perterra</i><c>per terra</c></SWT>  used in combination of W in third position:  · sei <RWW><i>sicura</i><c>al sicuro</c></RWW> adesso. |
| **X = auXiliary** | <https://universaldependencies.org/u/pos/AUX_.html>  · <SX><i>a</i><c>ha</c></SX>  · <SX><i>e</i><c>è</c></SX>  · <MX><i></i><c>ho</c></MX> detto |

**1.3 Third Position**

The third letter encodes the information about the grammatical categories involved and other phenomena.

|  |  |  |
| --- | --- | --- |
| 1 | A = aspect | used only with I in first position:  · <IVA><i>ha letto</i><c>stava leggendo</c></IVA> (solo perifrasi progressiva) |
| 2 | B = co-occurring tense and mood | 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> |
| 3 | G = gender | 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> |
| 4 | L = calque or linguistic transfer | 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> |
| 5 | M = mood | used only with I in first position:  · <IVM><i>aveva</i><c>avesse</c></IVM>  · <IVB><i>mettere</i><c>messo</c></IVB> |
| 6 | N = number | 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> |
| 7 | O = collocation | · cassetta degli <RNO><i>arnesi</i><c>attrezzi</c></RNO> |
| 8 | P = person | used only with I in first position:  · <IVP><i>esci</i><c>esce</c></IVP> |
| 9 | O = relative clause | used only with I in first position:  · il ladro <IVO><i>rubando</i><c>che sta rubando</c></IVO> |
| 10 | S = capitalization | 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> |
| 11 | T = Tense or Tokenization | used with I in first position means **TENSE**:  · <IVT><i>l'ha saputo</i><c>lo sapeva</c></IVT> (not IVA, although aspect is involved)  used with S in first position 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) |
| 12 | W = multi-Word expression | · <FWW><i>una borsa di patatini</i><c>un sacco di patate</c></FWW> |
| 13 | X = existential construction | · <SVScascade>[[1]](#footnote-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 |

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. How to use Transcript’omatic**

• open the TH file in the first editor, the LS in the second (below) and modify only the first one;

• in the boxes to the right of the text area (below “apri” and “chiudi”) write the tag (you can write also lowercase, it will be automatically changed in uppercase), for example type "sim" twice (the first time in the box below “apri” to create the opening tag and the second time in the box below “chiudi” to create the closing tag), you will read in the target hypothesis (TH) text area the tag "<SIM></SIM>";

• then write the incorrect and correct tags by clicking on the incorrect (left to write the opening tag and right to write the closing tag) and correct (again first click on the left and then on the right button) buttons; in this way, you will obtain "<SIM><i></i><c></c></SIM>";

• then move the cursor inside the tag <i></i> to write the incorrect LS word and write the correct TH word inside the tag <c></c> (e.g. <SIM><i>perche</i><c>perché</c></SIM>);

**3. 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, diphthongs and double consonants, we mark both the errors.[[2]](#footnote-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]](#footnote-3).

**3.1 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>**

**3.2 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 (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).

Immagine che contiene testo

Descrizione generata automaticamente

**We treat as tense errors those involving a missing auxiliary, e.g. mangio** 🡪 **ho mangiato, IVT no MX; but we mark MX in cases like this: visto 🡪 ho visto**

1. See CASCADE ERROR [↑](#footnote-ref-1)
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. [↑](#footnote-ref-2)
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>. [↑](#footnote-ref-3)