Annotating Second Language in Universal Dependencies: a Review of Current Practices and Directions for Harmonized Guidelines

Arianna Masciolini, Aleksandrs Berdicevskis, Maria Irena Szawerna, Elena Volodina
Språkbanken Text, SFS, University of Gothenburg, Sweden
{arianna.masciolini,aleksandrs.berdicevskis,maria.szawerna,elena.volodina}@gu.se

Abstract

Universal Dependencies (UD) is gaining popularity as an annotation standard for second language (L2) material. Grammatical errors and other interlanguage phenomena, however, pose significant challenges that official guidelines only address in part. In this paper, we give an overview of current annotation practices and provide some suggestions for harmonizing guidelines for learner corpora.

1 Introduction

Ever since Lee et al. (2017b) proposed to represent learner corpora as parallel dependency treebanks, Universal Dependencies (UD) (de Marneffe et al., 2021) has been gaining popularity as an annotation standard for both written and spoken second language (L2) data. At the time of writing, treebanks have already been released for written Chinese (Lee et al., 2017a), English (Berzak et al., 2016, now retired[1]), Italian (Di Nuovo et al., 2019, 2022) and Korean (Sung and Shin, 2024, 2025), as well as spoken English (Kyle et al., 2022). Two more are in progress, one for written Russian (Rozovskaya, 2024) and one for written Swedish.

When it comes to the annotation of L2 productions, the main advantage of using UD is that it provides a cross-lingually consistent morphosyntactic annotation layer that enables both quantitative and qualitative comparisons between a learner's L1 and L2, between different L2s and, most importantly, between the Target Language (TL) in its standard form and as an L2. In the latter scenario, an especially helpful format is that of parallel, so-called "L1-L2" treebanks, proposed by Lee et al. (2017b) and adopted by the vast majority of the seven aforementioned annotation efforts. In an L1-L2 treebank, each learner sentence is paired with a *correction hypothesis*, i.e. a corrected version of the learner's production based on an expert's interpretation of its intended meaning (if the sentence already adheres to the standard for the TL, the correction hypothesis is identical to the original sentence). As demonstrated by Masciolini (2023), this makes it possible to retrieve and analyze grammatical errors and other divergences between original learner sentences and their corrections via tree queries, without relying on any explicit error labeling scheme.

UD annotation of L2 productions, however, comes with its challenges. Learner texts often deviate from standard language in ways that make them difficult to analyze in a framework designed with standard use of the TL in mind. Relevant phenomena are not limited to grammatical errors themselves, but also include code switching, calques and non-idiomatic expressions. As a result, treebank developers have produced extensive project-specific guidelines, not always consistent with each other.

This study has been carried out as part of an ongoing annotation effort whose aim is to produce a treebank based on the SweLL-gold (Swedish Learner Language) corpus (Volodina et al., 2019), expanding the resources available through the Språkbanken research infrastructure. To ensure that the UD annotation of SweLL is consistent with that of other similar datasets, convenient for treebank users and theoretically motivated, we have reviewed the relevant UD guidelines and compared practices across five L2 treebanks for as many TLs. In this paper, we present our findings and propose directions for harmonizing L2 annotation guidelines. We demonstrate the application of our suggested principles on selected examples from our yet-to-be-released treebank.

2 Challenges of Treebanking L2

As mentioned in the introduction, a number of phenomena typical of — but not exclusive to — learner

[1] "Retired" UD treebanks are available for download but not up-to-date with current annotation guidelines and therefore not part of the latest UD release.

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 153–163
August 27, 2025 ©2025 Association for Computational Linguistics

language pose significant challenges for treebank developers. Most prominently, L2 productions are often characterized by the presence of grammatical errors, here understood in a broad sense, i.e. also including issues of orthography and lexical choice. Some of these issues, such as misspellings, are to some extent addressed in the universal UD guidelines (cf. Section 3.1) and do not typically alter the morphosyntactic analysis of the sentences in which they occur. However, it is not uncommon for grammatical errors to create a conflict between observed language use and intended meaning.

As a consequence, much of the discussion that has taken place in the context of L2 treebank development has revolved around trying to strike a balance between two seemingly contradictory approaches: following the principle of *literal reading* (as formulated by Berzak et al. 2016), according to which morphosyntactic analysis should be guided solely by the surface word forms and observed language use, and applying *distributional criteria*, in an attempt to reflect the writer's intentions. While annotating literally is often key to an informative analysis of nonstandard language, some amount of distributional insight is not only indispensable to deal with nonexisting word forms and otherwise unanalyzable constructions, but also intrinsic to the UD annotation scheme at large, independent of what material is analyzed.

We reframe the problem, guided by a similar but not identical question, i.e. to what extent the analysis of a learner sentence should be informed by its assumed intended meaning. While the principle of literal reading encourages not to rely on any particular interpretation of learner productions, it can be difficult to analyze ungrammatical constructions without some degree of speculation as to what the writer is trying to convey. In addition, when a correction hypothesis is available – such as in L1-L2 treebanks – it seems reasonable to try to ensure that the two analyses are consistent with each other, at least for the sake of comparability. Intuitively, the two trees constituting a sentence-correction pair should be "as different as necessary," so as to not hide any information that may help describe any discrepancies more accurately, and "as similar as possible" to facilitate fine-grained comparisons and error retrieval. An example showing the difference between literal and correction-driven annotation is given in Figure 1.

Another challenging situation, also illustrated in Figure 1, occurs when what the most informative annotation results in a category combination invalid in the underlying framework, which is descriptive of the standard use of the TL. This happens not only because of grammatical errors, but also in conjunction with other interlanguage phenomena such as code switching and syntactic calques. Relevant UD guidelines exist for some of these cases and will be discussed in Section 3.2.

3 L2 Phenomena in the Universal Guidelines

In this section, we summarize the general UD guidelines that relate to different aspects of the annotation of learner language. We restrict this discussion to the official guidelines available on the UD website,[2] although several proposals for unified guidelines for spoken data (Dobrovoljc, 2022) and user-generated content (Sanguinetti et al., 2020, 2023) also cover a variety of relevant phenomena.

3.1 Grammatical Errors

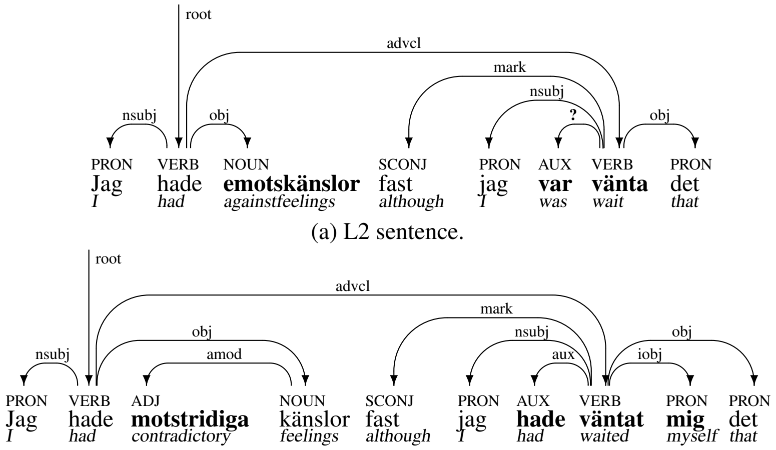
The universal UD guidelines comprise a page entirely dedicated to how to annotate typos, which has gradually evolved to also cover various types of strictly grammatical errors.[3] While it is explicitly stated that the recommendations are intended for dealing with sporadic errors rather than annotating learner corpora, they can be seen as a starting point and, as we will show in Section 4, they have been partially adopted by some of the existing treebanks.

When it comes to minor errors affecting individual tokens, the general idea is to never alter the word form itself, but signal the presence of an error with the feature `Typo=Yes`. Lemmatization should be based on the normalized spelling of the word, but morphological features should always describe the observed form. Additional information, such as the `CorrectForm` and features of the token, may be provided in the `MISC` field, creating a situation in which lemmatization is correction-driven, while morphological tagging follows the literal reading of the word.

This approach is easy to implement for simple **misspellings** and some clear-cut cases of **inappropriate lexical choice**, **incorrect morphological derivation** and **inflection**, but leaves open a wide range of problems, such as the analysis of unrecognizable forms and the syntactic annotation of words

[2] universaldependencies.org/guidelines.html

[3] universaldependencies.org/u/overview/typos.html



(a) L2 sentence.

(b) Correction hypothesis.

Figure 1: Example sentence-correction pair from SweLL. The correction hypothesis suggests that intended meaning is *I had mixed feelings although I had expected that.* The learner, however, uses the past tense of the copular verb *vara* ("to be"), *var* instead of *hade*, the corresponding form of the temporal auxiliary *ha* ("to have"). A **literal** reading of the verb group *var vänta* would use the dependency label cop. Conversely, a **correction-driven** approach would follow the correction hypothesis and opt for the relation aux – as would happen for the English construction *I was expecting that* – even though the aux-*vara* combination goes against the validation rules for Swedish.

that are incorrectly inflected e.g. for case – which is a marker for their syntactic role – or noncanonical in terms of their POS. All of these issues, however, have been addressed in at least some of the existing L2 treebanks.

When it comes to ***hypersegmentation***, i.e. incorrectly split words, annotators are essentially redirected to the well-established guidelines for the dependency relation goeswith, to be used in conjunction with the UPOS tag X and the Typo feature. For missing spaces, also referred to as ***hyposegmentation***, on the other hand, the recommendation is to handle them with the SpaceAfter=No and CorrectSpaceAfter=Yes attributes in MISC. **Missing words** are to be treated as ellipsis, whereas for **redundant tokens** it is suggested to follow the guidelines for speech disfluencies, thus using the reparandum relation. Issues arising from other **syntactic errors** are acknowledged by the existing guidelines, but not regulated in any way.

3.2 Foreign Expressions and Code Switching

UD provides guidelines for both treebanking extensively code-switched corpora and for annotating sporadic foreign words and expressions in otherwise monolingual treebanks.[4] In the latter case, more relevant for L2 corpora, the annotator may choose between three types of analysis: code-switched, borrowed and foreign.

If the treebanker opts for a *code-switched analysis*, content in a language other than the treebank's primary one is analyzed according to the language-specific guidelines of the former. To allow proper automatic validation, its ISO code is specified in the Lang attribute of the MISC field of each foreign token. Similarly, their FEATS field should contain the feature Foreign=Yes.

With a *borrowed analysis*, foreign words and expressions are assumed to have become part of the vocabulary of the main language of the treebank: they should not be marked as Foreign nor mandatorily assigned a Language, although the donor language may be specified through the OrigLang attribute. If the borrowed expression consists of several words, it is reduced to a flat structure and its tokens are POS-tagged according to the syntactic role of the expression as a whole.

Finally, a *foreign analysis* implies leaving the third-party language content completely unanalyzed: all relevant tokens are POS-tagged as X, marked as Foreign, optionally assigned an OrigLang and linked toghether using the dependency label flat, optionally subtyped as flat:foreign.

4 Annotation Practices Across L2 Treebanks

In this section, we provide a comparative analysis of the annotation practices found in different L2 treebanks. Its objective is twofold: on the one

[4] universaldependencies.org/foreign.html

Language	Name	Parallel	Modality	Sentences	Status	Publications	Annotation manual
Chinese	CFL	(✓)	written	451	released[5]	Lee et al. (2017a)	✓[a]
English	ESL	✓	written	5124	retired	Berzak et al. (2016)	
English	ESLSpok		spoken	2320	released	Kyle et al. (2022)	✓[b]
Italian	Valico	✓	written	398	released	Di Nuovo et al. (2019, 2022)	✓[c]
Korean	KSL		written	7530	released	Sung and Shin (2024, 2025)	
Russian		✓	written	500	in progress	Rozovskaya (2024)	
Swedish	SweLL	✓	written	~5000	in progress	this paper	

Table 1: Overview of UD treebanks of learner language.
[a] `github.com/UniversalDependencies/UD_CFL`; [b] `kristopherkyle.github.io/L2-Annotation-Project/dep_anno_overview.html`; [c] `github.com/ElisaDiNuovo/VALICO-UD_guidelines`

hand, to see to what extent the official guidelines discussed in Section 3 have been adopted in learner corpora; on the other, to investigate whether the annotation of similar phenomena is consistent across projects and what aspects of L2 annotation would benefit from harmonization. Our findings are summarized in Table 2.

To achieve these goals, we systematically review the papers describing the various treebanks, focusing on the sections dedicated to annotation criteria. When available, we integrate this with information with instructions from the treebanks' annotation manuals and complement textual sources with tree queries performed with the STUND tool for parallel treebanks (Masciolini and Tóth, 2024).

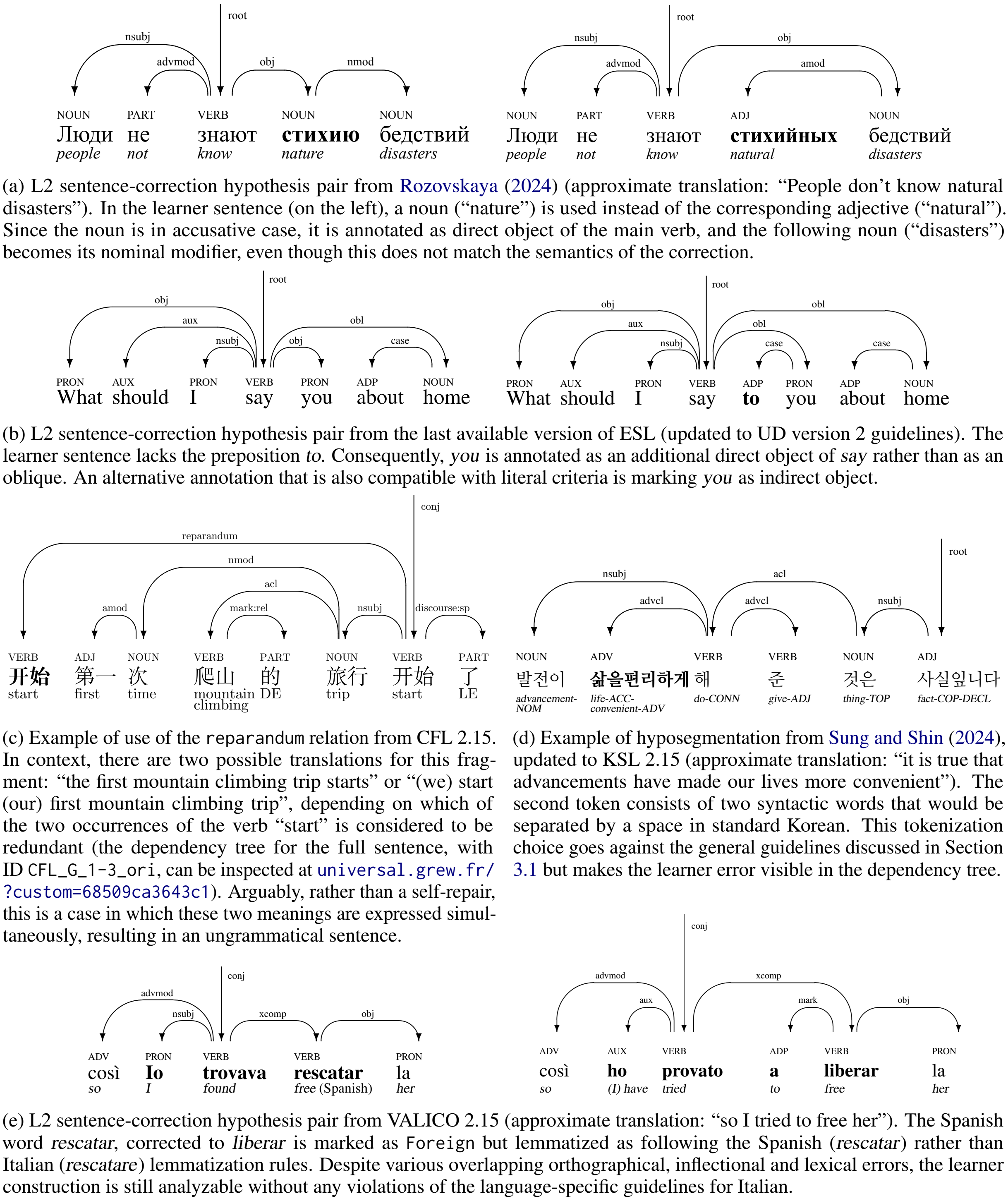
We consider five of the seven annotation efforts mentioned in the introduction and listed in Table 1: the Chinese as a Foreign Language treebank, henceforth **CFL** (Lee et al., 2017a);[5] the English as a Second Language treebank or **ESL** (Berzak et al., 2016); **VALICO**, based on the VALICO corpus of learner Italian (Di Nuovo et al., 2019, 2022), the Korean as a Second Language treebank, **KSL** (Sung and Shin, 2024, 2025) and the **Russian** learner treebank presented in Rozovskaya (2024), not yet released. We exclude the spoken L2 English treebank, ESLSpok, since neither the paper (Kyle et al., 2022) nor its annotation manual focus on L2-specific issues. The insights resulting from this comparison inform the guidelines for our upcoming SweLL-based treebank.

It must be noted that, for some annotation layers, CFL also provides a separate CoNLL-X file with double (literal and distributional) tags. In this paper, we focus on the annotation strategies that were chosen as the "default", i.e. on the values present in the CoNNL-U file distributed as part of the official UD releases. As for ESL, an important caveat is that the version described by Berzak et al. (2016) follows UD version 1 guidelines, which are now largely outdated. Furthermore, it should be kept in mind that such treebank is not annotated for `FEATS` or `MISC`. Similarly, CSL and KSL do not use the `FEATS` fields. Finally, the papers describing the Russian and Korean treebanks only discuss dependency annotation.

4.1 Grammatical Errors

Spelling Errors Misspellings of individual tokens only pose a problem in terms of lemmatization, unless they result in an ambiguous or completely unrecognizable word form. For uncomplicated cases, CFL and KSL, two out of three treebanks providing lemmas, follow the correction-driven approach suggested by the universal guidelines, lemmatizing based on the normalized form. While initially following the same principle, on the other hand, later versions of VALICO adopt a prevalently literal approach to the annotation of misspellings, trying to preserve the errors in the lemmas. When it comes to closed-class words, however, some exceptions are necessary to avoid validation errors. Misspelled auxiliaries, for example, are always assigned normalized lemmas since the language-specific guidelines for Italian only allow a restricted set of lemmas to be associated with the `UPOS` tag `AUX`. In cases of ambiguity or unrecognizable forms, KSL attaches the problematic token to the following one using the dependency label `flat`. None of the other treebanks appears to apply a similar strategy. No treebank, with the possible exception of Russian, uses the feature `Typo` or indicates the `CorrectForm` in the `MISC` field for errors of this kind, even though VALICO does use the former in cases of hypersegmentation (see below).

[5] The Chinese treebank was originally designed as a parallel treebank, but only its L2 half has been released.


root
nsubj
advmod
obj
nmod
NOUN PART VERB NOUN NOUN
Люди не знают стихию бедствий
people not know nature disasters
root
nsubj
advmod
obj
amod
NOUN PART VERB ADJ NOUN
Люди не знают стихийных бедствий
people not know natural disasters
root
obj
aux
nsubj
obl
obj
case
PRON AUX PRON VERB PRON ADP NOUN
What should I say you about home
root
obj
aux
nsubj
obl
obl
case
case
PRON AUX PRON VERB ADP PRON ADP NOUN
What should I say to you about home
conj
reparandum
nmod
acl
amod
mark:rel
nsubj
discourse:sp
VERB ADJ NOUN VERB PART NOUN VERB PART
开始 第一 次 爬山 的 旅行 开始 了
start first time mountain climbing DE trip start LE
root
nsubj
acl
advcl
advcl
nsubj
NOUN ADV VERB VERB NOUN ADJ
발전이 삶을편리하게 해 준 것은 사실입니다
advancement-NOM life-ACC-convenient-ADV do-CONN give-ADJ thing-TOP fact-COP-DECL
conj
advmod
nsubj
xcomp
obj
ADV PRON VERB VERB PRON
così Io trovava rescatar la
so I found free (Spanish) her
conj
advmod
aux
xcomp
mark
obj
ADV AUX VERB ADP VERB PRON
così ho provato a liberar la
so (I) have tried to free her

(a) L2 sentence-correction hypothesis pair from Rozovskaya (2024) (approximate translation: "People don't know natural disasters"). In the learner sentence (on the left), a noun ("nature") is used instead of the corresponding adjective ("natural"). Since the noun is in accusative case, it is annotated as direct object of the main verb, and the following noun ("disasters") becomes its nominal modifier, even though this does not match the semantics of the correction.

(b) L2 sentence-correction hypothesis pair from the last available version of ESL (updated to UD version 2 guidelines). The learner sentence lacks the preposition *to*. Consequently, *you* is annotated as an additional direct object of *say* rather than as an oblique. An alternative annotation that is also compatible with literal criteria is marking *you* as indirect object.

(c) Example of use of the reparandum relation from CFL 2.15. In context, there are two possible translations for this fragment: "the first mountain climbing trip starts" or "(we) start (our) first mountain climbing trip", depending on which of the two occurrences of the verb "start" is considered to be redundant (the dependency tree for the full sentence, with ID CFL_G_1-3_ori, can be inspected at universal.grew.fr/?custom=68509ca3643c1). Arguably, rather than a self-repair, this is a case in which these two meanings are expressed simultaneously, resulting in an ungrammatical sentence.

(d) Example of hyposegmentation from Sung and Shin (2024), updated to KSL 2.15 (approximate translation: "it is true that advancements have made our lives more convenient"). The second token consists of two syntactic words that would be separated by a space in standard Korean. This tokenization choice goes against the general guidelines discussed in Section 3.1 but makes the learner error visible in the dependency tree.

(e) L2 sentence-correction hypothesis pair from VALICO 2.15 (approximate translation: "so I tried to free her"). The Spanish word *rescatar*, corrected to *liberar* is marked as Foreign but lemmatized as following the Spanish (*rescatar*) rather than Italian (*rescatare*) lemmatization rules. Despite various overlapping orthographical, inflectional and lexical errors, the learner construction is still analyzable without any violations of the language-specific guidelines for Italian.

Figure 2: Selection of annotated sentence fragments from the existing L2 treebanks.

Incorrect Lexical Choices In the majority of cases, lexical choices are irrelevant to syntactic structure. However, problems can arise in at least two often overlapping cases: when the incorrect token is a function word and when its POS differs from that of the corresponding correct lexical item. Auxiliaries, for instance, are a closed class. If a lexical verb is used in place of an auxiliary, the annotator must therefore classify it as VERB, which in turn affects the dependency structure of the sentence. None of the available guidelines discusses this specific issue, but Di Nuovo et al. (2022) mention choosing the UPOS and DEPREL of misused closed-class words based on their literal reading. CFL makes abundant use of the generic dep relation type to solve cases in which the literally assigned POS tag is incompatible with any meaningful dependency label, whereas both ESL and the Russian treebank try to avoid it. At least in the case of ESL, this produces some violations of the current validation rules. On the opposite side of the spectrum, VALICO assigns POS tags according to the assumed intended meaning of the sentence.

Incorrect Derivation Issues related to derivational morphology can cause learners to inadvertently produce word forms that belong to a syntactic category other than intended one (cf. Figure 2a). Across treebanks, these cases are handled in the same way as incorrect lexical choices causing a change in POS (see above).

Incorrect Inflection The languages considered in this paper vary widely in terms of richness of their inflectional morphology. The general UD guidelines for Korean, for instance, do not discuss morphological features at all, while the Chinese treebank use them very sparsely. For this reason, neither the description of KSL nor the guidelines for CFL mention any issues of inflectional morphology. In English and Italian, the latter plays a more important role. In ESL, however, this is not discussed in depth since the treebank does not use universal morphological features.[6] In VALICO, on the other hand, inflectional errors are more prominent. Morphological features are assigned based on the observed word form, in accordance with the universal guidelines discussed above, and in the rare cases in which incorrect inflection can be seen as to alter the syntactic structure of the sentence, dependency annotation is consistent with the features. In Russian, the language with the richest morphology out of the ones addressed in this paper, the latter phenomenon is more frequent since the syntactic role of nominals with respect to their heads is marked by their case inflection. Consistently with Berzak et al. (2016) and Di Nuovo et al. (2022), Rozovskaya (2024) always selects the DEPREL of nominals based on their morphological suffixes, i.e. according to Berzak et al. (2016)'s definition, following literal criteria. A related phenomenon occurs in Korean, where particles indicate the syntactic role of the words they are attached to. In such cases, KSL guidelines suggest an approach where dependency annotation is guided by the intended use of the affected word, inferred from the context.

Hyper- and Hyposegmentation When it comes to hypersegmentation, the universal guidelines for the UPOS and DEPREL fields are generally followed, resulting in a widespread use of goeswith and X. For Korean, however, Sung and Shin (2024) explicitly mention one exception: when an extra space is found between a content word and its subsequent particle(s), the latter are treated as dependents of the former and the relation between them is labelled as case. In addition, ESL presents some exceptions when it comes to POS tagging: although the treebank is not entirely consistent in this respect, the tendency is to assign standard UPOS tags individually to each half of an incorrectly split word when the two segments are two complete recognizable words. In this sense, the most frequent case is that of incorrectly split compounds. None of the treebanks follows the universal guidelines when it comes to the MISC field. As for hyposegmentation, the only two treebanks for which the issue is explicitly discussed adopt two opposite approaches: VALICO follows the universal guidelines, thus splitting the incorrectly glued tokens and annotating them with SpaceAfter=No and CorrectSpaceAfter=Yes, whereas KSL leaves the tokens fused together, using the final morpheme as a cue to determine the DEPREL (cf. Figure 2d).

Missing Words At least to some extent, all five treebanks treat missing words as ellipsis. This works well for handling dependents lacking a head. However, missing function words – which are typically leaves – can alter the literal reading of a sentence in terms of syntactic structure. As a consequence, all treebank-specific annotation guidelines excepts CFL's mention a few cases that require special treatment. The most widespread issue is that of missing prepositions before non-core verb arguments, mentioned by both Berzak et al. (2016), Di Nuovo et al. (2022) and Rozovskaya (2024). VALICO seems to adopt a case-by-case approach, while in the English and Russian treebanks, these cases are consistently treated like incorrect case inflection in Russian: if an oblique is not introduced by a preposition, for example, it automatically becomes an obj or iobj (cf. Figure 2b).[7] Berzak et al. (2016) consider this to be one of the prime examples of literal – as opposed to distributional – annotation. We, however, argue that the use of

[6] XPOS tags, however, are assigned literally.

[7] As mentioned above, the version of ESL described in Berzak et al. (2016) actually follows UD version 1 guidelines, where direct objects were labelled dobj and there was no distinction between nmod and obl.

L2 Phenomenon	Guidelines	CFL	ESL	VALICO	KSL	ru
incorrect words	correct `LEMMA`	✓	–	✗*	✓	
	`Typo=Yes`	–	–	✗	–	
hypersegmentation	`goeswith`	–	✓	✓	✓*	
	`UPOS=X`	–	✓*	✓	✓	
	`Typo=Yes`	–	–	✓	–	
hyposegmentation	re-segmentation	–		✓	✗	
	`SpaceAfter=No`	–	–	✓	✗	
missing words	treat as ellipsis		✓	✓		✓
redundant words	`reparandum`	✓				✗
foreign material	(type of analysis)		`F`	`CS`		

(a) Adoption of universal guidelines for different macro-categories of L2 phenomena. A checkmark (✓) means that the treebank follows the general guidelines, whereas a cross (✗) implies the application of project-specific criteria. `F` and `CS` are abbreviation for, respectively, foreign and code-switched analysis. The `MISC` field is not considered here.

Error type	CFL	ESL	VALICO	KSL	ru
spelling errors	✓	✓	✗*	✓	
incorrect POS (derivational or lexical error)	✗	✗	✓		✗
incorrect `Case` inflection or particles	–	✗	✗*	✓	✗
hypersegmentation	–	✓*	✓	✓*	
hyposegmentation	–		✓	✗	
missing function words		✗	✓✗	✓	✗
redundant function words	✗	✗	✗	✗	✗
redundant content words	✓				
missing content words	✗	✗	✗	✗	✗
incorrect word order	✗				

(b) Application of correction-aware approaches to the annotation of different error types. A checkmark (✓) means that the annotation is at least partially correction-driven, whereas a cross (✗) implies the application of purely literal criteria.

Table 2: Overview of annotation practices across L2 treebanks. "–" signifies that the phenomenon and/or CoNNL-U field at hand are irrelevant for the treebank and/or language in question. The presence of any exceptions is signaled by asterisks (*). Cells for phenomena that are relevant but undocumented and/or unattested in the data are left blank.

`obl` and `obj`/`iobj` in this context are the results of two different distributional readings, both based on the context in which the oblique occurs, but only informed by the correction hypothesis in the former case. Similar problems arise for missing clitics (Italian) and particles (Korean). In VALICO, this can alter the dependency tree structure, whereas in KSL content words unmarked by the relevant particle are annotated as if the particle was present.

Redundant Words Across treebanks, the `reparandum` relation, whose use the universal guidelines encourage for redundant words, is attested five times, one in CFL and four in ESL. Only in the first case, however, its usage seems referred to a syntactically redundant word (cf. Figure 2c), while all ESL examples are akin to disfluencies in speech corpora. Guidelines for CFL, ESL, VALICO and the Russian treebank, however, mention a number of interesting subcases. In the English, Italian and Russian data, there are instances of redundant prepositions – a complementary problem to missing prepositions. In all three treebanks, the redundant preposition determines the syntactic role of the nominal it introduces, resulting in either `obl` and `nmod`. Furthermore, there are mentions of redundant clitics (Di Nuovo et al., 2019), aspects markers (Lee et al., 2017a), connectors and markers (Rozovskaya, 2024), all treated literally. Only Di Nuovo et al. (2022) mentions unrecognizable words, annotated with the `UPOS` tag `X` and the dependency label `dep`.

Other Syntactic Errors Other syntactic errors, such as atypical word order, are often not discussed at all in the treebanks' documentation. This is either because they do not affect the analysis of the sentence – which is the case in languages with no strict word order, such as Russian – or because their sparsity and heterogeneity encourages case-by-case decisions. Only the annotation guidelines for Chinese contain a paragraph that specifies how to handle words that appear misplaced: they should be attached to the closest possible head and marked as `dep`, i.e. left unanalyzed.

4.2 Other Interlanguage Phenomena

Surprisingly, only the VALICO documentation mentions interlanguage phenomena other than errors. According to Di Nuovo et al. (2019), **foreign words and expressions** are marked as Foreign and lemmatized based on standard rules for Italian. Later on, however, the project switched to a more clearly code-switched analysis using source language lemmas, although their language is not specified in the treebank (Di Nuovo et al., 2022). This, however, does not cause any validation issues since there are no foreign multiwords whose syntax breaks the language-specific validation rules for Italian (cf. Figure 2e). As for the other treebanks, only the feature Foreign is attested, exclusively in ESL and paired with the dependency label flat and with the POS tag X in all but one case, suggesting an approach similar to the foreign analysis in the universal guidelines.

Di Nuovo et al. (2022) also discuss ***syntactic calques***, i.e. phrases from the learner's L1 (or, more generally, from their extra-L2 linguistic repertoire) translated word-for-word to the target language. Arguing that this is the better choice in terms of cross-linguistic comparability, VALICO opts for analyzing the resulting expressions as if they were code-switched material in the source language, unless doing so results in a structure not valid in the Italian-specific UD guidelines.

5 Directions for Harmonized Guidelines

As shown in Table 2, neither the adoption of universal annotation guidelines relevant for L2 phenomena nor the application of correction-aware approaches to the annotation of learner errors is uniform across treebanks. Based on this comparative analysis and on insights from our first-hand experience treebanking SweLL, we propose a set of core principles for the harmonization of annotation guidelines for learner material.

In the following, we distinguish *token-level annotation*, i.e. the assignment of lemmas, morphological features and POS tags to individual words, from *syntactic annotation*, consisting in establishing and labeling head-dependent relations between pairs of tokens. We argue that the two can be guided by different, complementary principles, and that this results in more informative analyses.

5.1 Literal Token-level Annotation

As reported in Section 3.1, the universal UD guidelines call for a literal approach to POS and morphological tagging, but require lemmas to be inferred from the corrected version of the word form at hand. These principles are adopted in most project-specific guidelines, which also seem to agree on the fact that FEATS and UPOS should be assigned based on the literal criteria in a broader sense. VALICO is exceptional in that even the content of the LEMMA field is determined on the basis of the observed word form, e.g. preserving misspellings. We argue that this approach sometimes results in more informative analyses of certain learner errors and should be recommended for all learner treebanks. The application of this principle to a sentence from SweLL is exemplified in Figure 3.

PRON	VERB	ADJ	NOUN
Jag	har	flera	**förslagor**
jag	ha	flera	**förslaga**
I	*have*	*several*	*proposals*

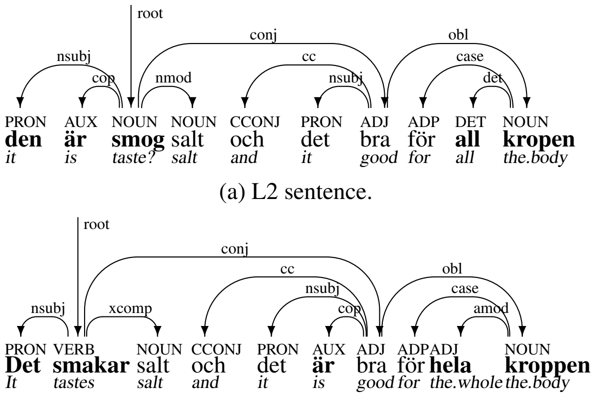
Figure 3: Sentence fragment from SweLL. The learner uses the incorrect plural form of the word *förslag*, inflecting it as if it belonged to a different declension. Lemmatizing as *förslaga* captures this phenomenon.

The Typo feature and the CorrectXXX attributes are not widely used, perhaps because they are considered redundant for parallel treebanks, where the corrected form, lemma and features of a token can be inferred from the correction hypothesis. Monolingual L2 treebanks, on the other hand, could benefit from their more systematic use. If lemmatization is performed as suggested above, the introduction of a CorrectLemma field would make the annotation more complete and could be checked by the UD validator instead of the LEMMA field itself to address potential lemma-POS mismatches such as the ones discussed in Section 4. We therefore encourage the use of such fields, which are important for monolingual L2 treebanks and can be largely filled in automatically in parallel datasets.

Finally, universal guidelines for hypersegmentation are widely adopted, but this is not the case for hyposegmentation. In the latter case, we strongly support KSL's choice not to re-segment incorrectly merged tokens, as doing so makes learner errors invisible to many UD tools.

5.2 Correction-aware Syntactic Annotation

As discussed in Section 2, we refrain from a discussion of syntactic analysis in terms of the two categories "literal" and "distributional" as we argue that dependency annotation is always, to some extent, distributional. Instead, we focus on the role of corrections in the annotation process and recommend a *correction-aware* approach. This is not to say that dependency relations used in an L2 sentence should follow those found in the annotation


root
conj
obl
nsubj
cc
case
cop
nmod
nsubj
det
PRON AUX NOUN NOUN CCONJ PRON ADJ ADP DET NOUN
den är smog salt och det bra för all kropen
it is taste? salt and it good for all the.body
root
conj
cc
obl
nsubj
case
nsubj
xcomp
cop
amod
PRON VERB NOUN CCONJ PRON AUX ADJ ADP ADJ NOUN
Det smakar salt och det är bra för hela kroppen
It tastes salt and it is good for the.whole the.body

(a) L2 sentence.

(b) Correction hypothesis.

Figure 4: Example sentence pair from SweLL. The L2 production is syntactically unclear, but the correction hypothesizes the intended meaning to be *It tastes salt and it is good for the whole body*. *smog* is therefore interpreted as a misspelling of *smak* ("taste", noun), in a construction that can be literally translated as *it is a taste (of) salt*. Furthermore, the adjective *bra* is considered to be the head of the second conjunct even if the original phrase lacks a copula. The resulting trees share a similar high-level structure but diverge enough to capture all grammatical errors.

of its correction hypothesis – which is in any case not available in non-parallel treebanks – but rather that the annotator should ground syntactic analysis in both the observed language use and the assumed intended meaning. We argue that this is the only informative way to analyze syntactically unclear sentences, which are a relatively common occurrence when the language use deviates from the standard for the TL significantly. In parallel treebanks, this also guarantees the analysis of learner sentences is consistent with that of the corresponding CHs, improving comparability.

Correction-aware syntactic annotation may appear contradictory with what we proposed in Section 5.1. Rather, we see these two principles as complementary, as we illustrate in Figure 4. As shown in Figure 1, however, there are cases in which applying literal criteria at the token level and determining the dependency structure of the sentence based on the distributional properties of the words results in category combinations not allowed by the UD validator. We therefore propose the use of a new relation subtype, `*`, to be used to mark deliberate violations of the UD universal and/or language-specific guidelines in correspondence of an L2 phenomenon that itself goes against the norms of the TL. An example of its intended usage is given in Figure 5.[8]

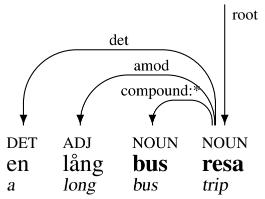


Figure 5: Sentence fragment from SweLL. In standard Swedish, compounds are written as single words (*bus trip*, for instance, translates to *bussresa*). The label `compound` should therefore only be used for borrowed compounds. Since it can be argued that the learner sentence mimics a construction from another language, we propose the use of the label `compound:*` and the POS tag `NOUN` for both *bus* and *resa* as a more informative alternative to `goeswith` and `X`.

5.3 Syntactic Calques as Code-switching

When it comes to foreign content, we recommend VALICO's approach, which allows for more informative annotations than ESL's foreign analysis. Taking inspiration from both VALICO's treatment of syntactic calques and ESL's approach to POS tagging incorrectly split compounds, we propose extending the recommendations for code-switched analysis even to constructions where TL words are combined into a construction reminding of the learner's extra-TL repertoire, with minor adaptations: while using the `Lang` feature may be misleading for tokens that, individually, correspond to words of the primary language of the treebank, the use of `*` would make it clear that the construction is nonstandard without strictly binding it to the validation rules of the calqued language. An example is given in Figure 5.

6 Concluding Remarks

We discussed the challenges L2 poses for UD annotators and gave an overview of the universal guidelines relevant for learner corpora, as well as of the current annotation practices across five of the existing treebanks. We concluded the paper with some suggestions for harmonization, which we hope can be the starting point for a productive discussion, primarily with other members of the UD community working directly with L2 material, but also with annotators of spoken and/or code-switched data, user-generated content and nonstandard language varieties such as pidgins and dialects, all of which share some features with learner language.

[8] Similarly, `*` could be used in the example shown in Figure 1 to allow analyzing *var* as an `aux`.

Acknowledgments

This work has been funded by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161) and supported by the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). In particular, the 2024 UniDive training school in Chișinău provided the first author with excellent opportunities for annotation experiments and preliminary discussions of L2-related issues with other members of the UD community. We also want to thank Xurui Zhang for his invaluable help with interpreting the annotation of the Chinese examples, as well as Dana Dannélls and the SyntaxFest 2025 reviewers for their thoughtful and supportive feedback at different stages of the writing process.

References

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Elisa Di Nuovo, Cristina Bosco, Alessandro Mazzei, and Manuela Sanguinetti. 2019. Towards an Italian learner treebank in Universal Dependencies. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR-WS.

Elisa Di Nuovo, Manuela Sanguinetti, Alessandro Mazzei, Elisa Corino, and Cristina Bosco. 2022. VALICO-UD: Treebanking an Italian learner corpus in Universal Dependencies. *IJCoL. Italian Journal of Computational Linguistics*, 8(8-1).

Kaja Dobrovoljc. 2022. Spoken language treebanks in Universal Dependencies: an overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. 2022. A dependency treebank of spoken second language English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45, Seattle, Washington. Association for Computational Linguistics.

John Lee, Herman Leung, and Keying Li. 2017a. Towards Universal Dependencies for learner Chinese. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 67–71, Gothenburg, Sweden. Association for Computational Linguistics.

John Lee, Keying Li, and Herman Leung. 2017b. L1-L2 parallel dependency treebank as learner corpus. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 44–49, Pisa, Italy. Association for Computational Linguistics.

Arianna Masciolini. 2023. A query engine for L1-L2 parallel dependency treebanks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 574–587, Tórshavn, Faroe Islands. University of Tartu Library.

Arianna Masciolini and Márton A Tóth. 2024. STUnD: ett Sökverktyg för Tvåspråkiga Universal Dependencies-trädbanker. In *Proceedings of the Huminfra Conference*, pages 95–109, Gothenburg, Sweden.

Alla Rozovskaya. 2024. Universal Dependencies for learner Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17112–17119, Torino, Italia. ELRA and ICCL.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020. Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5240–5250, Marseille, France. European Language Resources Association.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2023. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. *Language Resources and Evaluation*, 57(2):493–544.

Hakyung Sung and Gyu-Ho Shin. 2024. Constructing a dependency treebank for second language learners of Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3747–3758, Torino, Italia. ELRA and ICCL.

Hakyung Sung and Gyu-Ho Shin. 2025. Second language Korean Universal Dependency treebank v1.2: Focus on data augmentation and annotation scheme refinement. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 13–19, Tallinn, Estonia. University of Tartu Library, Estonia.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and 1 others. 2019. The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology*, 6:67–104.