An Annotated Error Corpus for Esperanto

Eckhard Bick
University of Southern Denmark & GrammarSoft ApS
eckhard.bick@gmail.com

Abstract

This paper presents and evaluates a new multi-genre error corpus for (written) Esperanto, *EspEraro*, building on both learner, news and internet data and covering both ordinary spelling errors and real-word errors such as grammatical and word choice errors. Because the corpus has been annotated not onlyse for errors, error types and corrections, but also with Constraint Grammar (CG) tags for part-of-speech, inflection, affixation, syntactic function, dependency and semantic class, it allows users to linguistically contextualize errors and to craft and test CG rules aiming at the recognition and/or correction of the various error types covered in the corpus. The resource was originally created for regression-testing a newly developed spell- and grammar checker, and contains about 75,000 tokens (~ 4,000 sentences), with 3,330 tokens annotated for one or more errors and a combined correction suggestion. We discuss the different error types and evaluate their weight in the corpus. Where relevant, we explain the role of Constraint Grammar (CG) in the identification and correction of the individual error types.

1 Introduction

Error corpora have the potential to play an important role in modern linguistics, and can support diverse tasks such as pedagogical-didactic work, the development of spell- and grammar checkers, as well as research on language change and variation. Most of these corpora, however, are error corpora in the specialized sense that they contain L2 learner data covering one or more L1 languages, e.g. (Al-Jarf, 2010) with an overview for English, (Rakhilina et al., 2016) for Russian, or (Arnardóttir et al., 2022) for Icelandic.

Gamon et al. (2013) stress the three-fold importance of such corpora for automatic systems: error statistics, ML training and evaluation. However, it remains unclear how well L2 data carry over to native speaker errors. In addition, many of the available L2 resources are limited by the fact that they do not provide an actual error mark-up, let alone systematic error classification. The European CLARIN initiative[1], for instance, offers 75 learner corpora covering 15 languages. Yet only seven of these corpora are listed as providing actual error labels[2]. Also, and not least in a world-wide perspective, English is by far the best-represented L2, followed by a few dozen major and European languages at most, and no data at all for the vast majority of languages. Thus, in a global overview of learner corpora at the University of Louvain-la-Neuve[3], out of 208 corpora, 50% have English as the target language, followed by Spanish, German and Italian with 9% each, French (6%), Chinese, Russian (2%) and Arabic (1%).

An example of an error corpus that is *not* a learner corpus is Sketch Engine's *Error Corpus from English Wikipedia,* with seven error types (spelling, lexico-semantic, typos relating to style, punctuation, typographical, other). However, error tagging is unrevised, automatic only and mostly focuses on safe typographical errors[4], not mentioning grammatical errors. An interesting alternative, providing no error classification, but at least a corpus of manual corrections, is to export edit histories from Wikipedia (or other sources), a method suggested by Lichtarte et al. (2019) to procure training data for the automatic

[1] https://www.clarin.eu/resource-families/L2-corpora
[2] for English, Czech, Norwegian, Slovene, Hungarian, Spanish, German/Italian
[3] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html
[4] https://www.sketchengine.eu/error-corpus-from-english-wikipedia/

correction of grammatical errors. On a general note, shared task training data is a valuable source of structured and comparable corpus data in the ML community, and the upcoming *Shared task on Multilingual Grammatical Error Correction* at the 2025 NLP4CALL workshop[5] will involve sentence correction pairs for at least 10 languages, albeit without error classification proper.

EspEraro, the Esperanto corpus presented here - to the best or our knowledge the first of its kind - is intended to fill a data gap for this under-resourced language, providing a wide scope of errors, with all data fully analyzed at various linguistic levels. Unlike many other error/learner corpora, EspEraro provides not just learner data or other error-containing material, but points out where and what the errors are, with a fine-grained error classification system for real-word errors, and manual revision for all mark-up.

2 The corpus

EspEraro contains about 4,400 Constraint Grammar-annotated sentences with 75,000 tokens, of which 3,330 are marked for one or more error types, as well as a combined correction suggestion. The CG tags used adhere to the cross-language VISL convention (Bick 2023). The corpus was developed as part of an ongoing spell- and grammar checker project Lingvohelpilo-2[6], in which it was used for regression testing[7], and the inclusion of error types and sentences in the corpus was often motivated by challenges encountered during development. Therefore, in addition to presenting the various error types and evaluating their relative impact on the corpus (section 3), we will also discuss the methods that were used to identify and correct them, with a special focus on the role of Constraint Grammar, explaining some of the about 2,000 rules used in the Lingvohelpilo-2 project.

Corpus material was added from three different sources, reflecting different phases of development:

(a) learner sentences from the online teaching portal Lernu![8], containing a non-systematic variety of errors both orthographical and grammatical

(b) text book error examples from various sources and modified dictionary quotes[9], systematically covering grammatical errors and certain particles

(c) corpus sentences containing lexical errors (word choice, false friends, confusable word pairs), plus random errors co-occurring with the former.

Type (c) material constitutes the main part of the collection and was harvested from a 200 million word reference corpus of news, literature, wikipedia and a variety of internet sources, hoping to capture not just learner errors, but a representative cross section of frequent errors made by the language community as a whole. In this sense, the corpus transcends what is normally understood as an L2 corpus. However, the roots of Esperanto are artificial, and the language is not linked to any regionally defined political entity, nation or ethnic group, and with the notable exception of mixed-marriage native speaker children, most language users are therefore L2 speakers. Given the language's regular grammar and affixation system, the learning period proper will be considerably shorter than for other languages, but even fluent speakers may still make errors, not least caused by L1 interference or first foreign language interference (often English). In addition, due to the lack of a dominant native speaker community, there is a high tolerance for lexical and syntactic variation, exerting less pressure on the individual to become aware of and avoid interference errors. Because of these factors, the borders between a learner-error corpus, an L2 corpus and a wider error corpus for "native" speakers is blurred in the case of Esperanto.

The reference corpus is available for searching and statistical analysis in the CorpusEye interface (https://corp.visl.dk), and contains morphosyntactic, dependency and semantic annotation provided by the EspGram Constraint Grammar parser (Bick 2007). We used this corpus in a dual fashion. First, to identify frequent orthographical errors and error patterns,

[5] https://spraakbanken.gu.se/en/research/themes/icall/nlp4call-workshop-series/nlp4call2025
[6] https://lingvohelpilo.visl.dk/, with financial support from ESF (Esperantic Studies Foundation)
[7] This is why the corpus also contains a certain amount of error-free sentences, covering cases where false positive markings where addressed. For the same reason, sentences are not contiguous, but collected.
[8] https://lernu.net
[9] The main dictionary sources were ReVo (https://reta-vortaro.de) and PIV (Plena Ilustrita Vortaro, http://vortaro.net

we ran statistics on lower-case words marked as heuristic or compounded and frequency-sorted the result for inspection and automatic look-up of their lemmas[10]. Obviously, the method also throws up foreign words, neologisms and unrecognized derivations, but it did help to identify frequent error patterns, e.g. confusion of consonant or vowel pairs, or jumbled consonant clusters, as well as OCR artifact error patterns (e.g. 'm' vs. 'rn' [r+n]) and Esperanto-specific encoding confusion issues with diacritics (e.g. ' ,' = ĝ, 'ÿ' = ŝ, '∑' = Ŝ , 'u˘ ' = ŭ etc.).

Second, we used the corpus to find usage examples of grammatical errors, for instance agreement clashes in NP's (number or case) or between subject and subject predicative (number). We also checked for transitivity errors by doing dependency searches for intransitive verbs having direct objects, or looking for passive participles of intransitive verbs. Esperanto does not allow dual transitivity usage and uses affixed to make intransitive verbs transitive or vice versa, so the method will flag either usage errors or affixation errors. Finally, we used the reference corpus to look up false friends or word confusion pairs suggested in the literature, identifying grammatical and lexical trigger contexts for the mapping rules marking the confusion error in question. At the same time, one or more typical error usage cases, and sometimes a correct counter example, would go into the EspEraro corpus for regression testing.

3 Error types

Currently, the error corpus contains 3860 individual instances of error type markings[11] attached to 3330 "annotation tokens"[12]. The error markings come in three variants, depending on the marking method used in the proofing tool. Non-word errors that are similar to an existing word xxx (a) are marked with the latter in the form of an <R:xxx> correction tag and a <W:[0-9]+> Levenshtein similarity value[13]. Errors found through lexical lookup and pattern-recognition (b) are classified as <E-TYPE:....>, and CG-mapped error types (c) use a CG-recognizable prefix (@....). Correction suggestions <R:...> for (b) and (c) are either generated on-the-fly in conjunction with error-classification, or post-generated based on corrected POS and inflection tags.

The CG-annotated example sentence in figure 1 contains — marked in red — a non-word *(konfrmis),* a lemma substitution error *(perfortigita),* two accusative case errors *(morganata[n] geedzeco[n])*, one transitivity error *(translokis)* and an article insertion *(la).* Depending on how the sentence is interpreted, there could also be a subclause-end comma before *'kaj'* (marked on the latter as *@comma-FSend).* Note that the annotation retains, in addition to the error tagging, all relevant VISL CG tag fields:

Word [lemma] <secondary> POS MORPH &syntax #dependency

with <secondary> containing tags for subclass (e.g. <rel> relative pronoun, <fem> = female, <mv> = main verb), valency (e.g. <vt> = transitive verb), semantic prototype (e.g. <Hprof> = human professional, <Ltown>), framenet (e.g. <fn:hurt>), sentiment (<Q+/->), derivation/compounding (e.g. <F:ge%edz%ec|o>) and coordination structure (e.g. <cjt-first>).

Ŝi [ŝi] <fem> PERS 3S F NOM &SUBJ> #1->2 *She*
anoncis [anonci] <vq> <fn:declare> <mv> V IMPF VFIN &FS-STA #2->0 *declared*
, [,] PU #3->2
ke [ke] KS &SUB #4->6 *that*
ŝi [ŝi] <fem> PERS 3S F NOM &SUBJ> #5->7 *she*
estis [esti] <aux> <cjt-first> V IMPF VFIN &FS-<ACC #6->2 *was*
perfortigita [devigi] **<R:devigita>** <fn:be_attribute> <Q-> <fn:hurt> <PRP:per+fort%ig|i> <mv> V PCP PAS IMPF ADJ S NOM &ICL-AUX< &ICL-AUX< **@:BASE-devigi** #7->6 *forced*
rezigni [rezigni] <fn:refrain> <Q-> <mv> V INF &ICL-<SA #8->7 *to resign*
kaj [kaj] <clb> KC &CO #9->6 *and*
translokis [transloki] **<R:translokiĝis>** <cjt> <PRP:trans+lok|i> <fn:transfer> <mv> V IMPF VFIN &FS-STA **@iĝ** #10->2 *moved*
al [al] PRP &<OA #11->10 *to*
Romo [Romo] <Ltown> PROP S NOM &P< #12->11
, [,] PU #13->12 *Rome*
kie [kie] <rel><aloc> ADV &ADVL> #14->21 *where*

[10] In Esperanto, lemmatization is not in itself an error source, because after filtering of a few function words, word class and inflection can be safely predicted from endings.
[11] not counting the secondary tag of "green" (not a serious error, or not regarded as an error by some).
[12] An annotation token may not necessarily be space-delimited. Thus, multi-part named entities are regarded as one token, and for word splitting and word fusion may be annotated together as one token. Also, word insertions are counted as tokens.
[13] Computed based on editing distance, phonetic similarity, keyboard likelihood and frequency.

la [la] ART **@insert** #15->15 *the*
tiama [tiama] <jtime> ADJ S NOM &>N #16->17
then pope Gregory the 16th
papo [papo] <Hprof> N S NOM &SUBJ> #17->21
Gregorio [Gregorio] PROP S NOM &N< #18->17
la [la] ART &>N #19->20
16-a [16-a] <num-ord> ADJ S NOM &N< #20->18
konfrmis [konfirmi] <vq> <fn:confirm>
<R:konfirmis> **<W:2124>** <mv> V IMPF VFIN
&FS-N< #21->12 *confirmed*
ŝian [ŝia] <fem> <poss> DET S ACC &>N #22->24
her
morganata [morganata] **<R:morganatan>** ADJ S
ACC &>N **@acc** #23->24 *morganatic*
geedzeco [geedzeco] **<R:geedzecon>** <F:ge%edz
%ec|o> <f-soc> N S ACC &<ACC **@acc** #24->21
. [.] PU #25->2 *marriage*

Figure 1: Annotated sentence

About 15.7% of all errors are examples of purely orthographical errors, or 13.3% if casing errors are excluded. About 40% of these[14] are similarity-based corrections of unrecognized words. In addition to Levenshtein similarity weighting (cf. <W:[0-9]+> tags), we rely on the ordinary CG disambiguation rules of the EspGram parser to weed out contextually unfit correction suggestions. Another 40% of the simple spelling errors are identified using a special dictionary of orthographical error patterns and deprecated word forms. The remainder (20%) is handled by a mixture of (lexicon-informed) preprocessing and CG rules. The former handles encoding-based errors and some OCR errors, alongside the normalisation process for Esperanto diacritics[15]. It may also suggest word fusions or splittings that will then be validated using CG rules. The latter can also add word fusions themselves, based on "impossible" context such as same-case N-N chains. In addition, CG is used to mark and correct hyphenation errors and numbering format errors, e.g. dates, ordinal endings and decimal markers, as well as marking upper case / lower case errors. Here, SUBSTITUTE rules with unifications variables are used to suggest corrections, while ADD rules are used for error tags that leave the actual correction to an external generator (e.g. @upper, @lower):

```
(r1) SUBSTITUTE ("[^<]+"r)
("$1-$2"v <error> <E-TYPE:missing-hyphen>)
TARGET ("([A-Z][A-Z]+|[0-9]+)([a-z]+)"r <heur>)

(r2) ADD (@lower) TARGET <*>
(0 <jnat> OR ("[a-z]+e"r <PROP:.*\\|e>r ADV))
(-1 ALL-ORD - <*>) (NOT 1 (<*>))
```

Rule (r1) marks and inserts a missing hyphen in words like *'UEA(-)delegito'* (UEA delegate) and *'3(-)ĉambra'* (3-bedroom). The simplified rule (r2) marks uppercase <*> nationality adjectives <jnat> and name- (PROP) derived adverbs (ADV) as lower case (@lower), unless they are sentence-initial or part of an uppercased MWE, i.e. with an upper-case word before or after.

The spellchecker offers the user a fine-grained choice of activating or inactivating individual error types, in order to avoid having to look at and ignore false positive markings. In the same vein, the tool marks dictionary-wise unknown names, compounds and derivations as such rather than throwing up an error[16]. Foreign words are much less safe to distinguish from spelling errors, so a tag is shown (@foreign, 2% of markings in the corpus). Words that are not likely to be either, and do not have a small edit distance to a known word, are marked as @new[17] (1.3% in the corpus). These may be true lexicon gaps, e.g. *trafikarpo* (a tree), *ortparalele,* name-like nouns, e.g. *laolumoj, tuĝja-oj* (ethnic groups) or unrecognized, mostly Romance, foreign words, e.g. *protezione, geofisica, piranha.*

In terms of Constraint Grammar, real word-errors, i.e. grammatical and word choice errors, are the most relevant categories, because they can only be handled by including context and semantics. Thus, rule (r3), mapping a direction-accusative ending on place nouns (N-LOC) relies heavily on both valency (e.g. for adverbial arguments <va+DIR/LOC>,<vta+DIR/LOC>),

[14] i.e. of the 13.3%
[15] Depending on keyboard options, the circumflex in the letters ĉ, ĝ, ĥ, ĵ, ŝ and ŭ is sometimes replaced with an 'x' or an 'h'. Whereas the former is almost always uniquely reversible, the latter may create a small amount of ambiguity to be resolved.
[16] EspEraro retains the full morphological analysis, so there will be a POS tag for names (PROP) and a segmented analysis for compound parts and affixes for complex words.
[17] These may be foreign words that *look* like Esperanto words (e.g. *vulcanologia*), esperantized Chinese roots (e.g. *ĝingluo-o*) or triple-compound words (e.g. *hom+riĉ+font%an,* with + being a root boundary and % an affix boundary)

semantic prototypes (e.g. <Lpath>, <con> [container], <an.*> [anatomicals], <cc-h> [made things]) and a semantic verb set (V-MOVE-I [intransitive movement verbs]):

```
(r3) ADD (@acc-dir) TARGET N-LOC + NOM
(*-1 PRP-LOC BARRIER NON-PRE-N/ADV LINK NOT
0 ("trans") OR ("ĉe") OR ("ĉirkaux") LINK *-1 VV
BARRIER NON-ADV OR <adir> LINK 0 <va+DIR> -
V-MOVE-I - <ve> OR <vta+DIR> LINK NOT 0
<va+LOC> OR <vta+LOC>)
(NOT 0 <Lpath> OR <an.*>r)
((NEGATE 0 <con> OR (<cc-h>) LINK p ("en"))
OR (*p ("meti") OR ("ŝuti") OR ("verŝi")))
```

With a narrow definition, leaving out word insertions, deletions and substitutions, there are about 35 grammatical error categories, covering POS, affixation and inflection errors, that amounts to 22.0% of all errors in the corpus.

Tag	**Error**	**%**
acc	accusative -n: object	31.6
acc-dir, acc-oc, -quant, -trans	accusative -n: other complements	14.1
nom, nom-oc, -pcp, -prp	nominative: all uses	15.6
refl, no-refl	reflexive pronoun	1.5
adj, adv, noun	POS errors	4.0
-io, -iu	correlative pronoun	0.9
pl	plural	8.9
sg	singular	3.8
akt, pas	active vs. passive	1.1
ata, ita	aspect	1.1
vfin, as, is, os	finity/tense	4.9
us, u	finity/non-tense	0.9
inf	infinitive	1.3
ig/iĝ (affixes)	missing (in)transitivity	5.9
DEL:ig/iĝ	spurious (in)transitivity	3.4
ado, ul	other affixes	0.9
		99.9

Table 1: Grammatical errors

As can be seen from table 1, Esperanto's only case inflection ending, the accusastive -n, accounts for almost half the errors (45.7%) when including both nominal (object/subject, argument of preposition), predicatives and andverbial functions (direction, quantity). Note that while the corpus consists of chosen examples and for the sake of coverage needs to over-represent small error classes, the high frequency of accusative errors is nevertheless indicative, if not representative, for the language as a whole, as these errors also co-occur as "by-catch" in many sentences chosen for the sake of other errors.

Number agreement errors (sg, pl) make up the second largest group (12.7%) followed by transitivity suffix (ig/iĝ) errors (10.2%) with 10.2% verb inflection errors (9.4%). As might be expected for categories where one alternative is unmarked, wrongly adding an inflection ending or suffix is less frequent than wrongly omitting it. Thus, using a plural ending is more "premeditated" and less likely to be wrong (sg) than forgetting it (pl), and using an explicit transitivity marker spuriously is less likely than simply using an intransitive verb root with an object or as a passive – a "false friend" usage that is seen in English, German and many other natural languages (e.g. *to begin, end, roll, run, open)*.

A few grammatical error categories, while tagged on one individual token, have a wider scope and imply changes in more than one token, in which case no <R:...> correction tag can be shown. Thus, the lack of a subject is marked on the finite verb (@PREADD:subject), 3.2%) and inverted order is marked on the second word (@PRESWAP, 1.4%). Finally, a @warning tag (0.6%) is used on tokens with a syntactically impossible context, but no simple/local correction. Finding this type of error is a CG task. (r4), for instance, throws a warning for finite verbs (VFIN) directly following left-attaching subjects (&<SUBJ) without a left context of a clause-boundary word (CLB-ORD) or a right-attaching sister subject (sl &SUBJ>):

```
(r4) ADD (@warning) TARGET VFIN
(-1C &<SUBJ LINK *-1 VFIN LINK NEGATE *-1 CLB-
ORD)
(NEGATE -1 &<SUBJ LINK sl &SUBJ>)
```

This rule calls for a 1-verb rephrasing of a sentence with two clashing verbs, likely originating from a copy-paste error:

Cetere *(by the way)*, la 23an de junio *(23 June)* **okazas** *(takes place)* la sporta mondo *(the sport world)* **celebras** *(celebrates)* ĉiujare *(every year)* "Olimpika Tago"-n *(an Olympic Day-acc)*

In addition to ordinary spelling errors and grammatical errors, we also mark semantic errors where the words' lemma itself, rather than its spelling, inflection or affixation is wrong. Lemma errors amount to 17.0% of all errors in the corpus and come in two types, local and contextual. The first are lemmas that look like Esperanto, but aren't – with entire lemmas or word parts inspired by other languages, e.g. *net|komunumo* (internet community) instead of *ret|*

komunumo, where the morpheme for "internet" is wrongly assumed to be 'net-', a root that in Esperanto only has the meaning of 'after costs' (as in *net profit*). Once known to the system, this kind of lemma error can be safely corrected with a dictionary look-up.

Contextual lemma errors, on the other hand. are arguably the most difficult type of real-word errors, comprising confusion errors, false friends and word choice errors based on a wrong meaning. In all of these, lemma substitution has to be performed with few or no surface clues other than valency patterns and the semantics of the surrounding words, which is why this kind of error can only be handled with complex and lemma-specific CG rules. Thus, (r5) lemma-corrects *'oferi'* (to sacrifice) into *'oferti'* (to offer) if it has an accusative (ACC) dependent (c) that is not a person (HUM-pers), animal (<A[a-z]*>), anatomical (<an.*>) or 'comfort':

(r5) ADD (@:BASE-oferti) TARGET ("oferi")+AKT
(c N/PROP/PRON + ACC LINK NOT 0 HUM-pers
OR(<A[a-z]*>r) OR (<an.*>r) OR ("komforto"))

Tag	Error	%
lemma	non-word lemma error	10.5
:BASE-xxx	real-word lemma confusion, inflecting	62.3
:xxx	real-word lemma confusion, non-inflecting	27.2
		100

Table 2: Lemma errors

Contextual lemma errors are 9x more frequent in the corpus than non-word lemma errors (table 2). About 70% of the former are inflecting lemmas and need postprocessing after substitution. The remainder are function word errors, mostly wrong use of a preposition, a common example being the confusion of *'de'* (*la aŭto de amiko* – the car of a friend) and the "quantitative" preposition *'da'* (*botelo da vino* – a bottle of wine).

Another contextual and CG-managed error type are insertions and deletions, with 5.7% and 2.6% of error markings, respectively. A typical rule is (r6), inserting the conjunction *'ke'* (that) after af finite verb form (VFIN) of '*pensi'* (think), if it has a left subject dependent (cl &SUBJ>) and is followed (*1) by another left subject (and then its verb) without interfering material other than prenominals (NON-PRE-N/ADV). The conjunction is obligatory in Esperanto and omitting it is a common "English" (or, for that matter, Scandinavian) L1 interference error:

(r6) ADDCOHORT
("<ke>" "ke" KS &SUB @insert)
AFTER ("pensi") + VFIN
(cl &SUBJ>)
(*1 &SUBJ> BARRIER NON-PRE-N/ADV OR CLB-ORD LINK *1 VFIN BARRIER CLB)

Insertions concern mainly missing definite articles (69%), followed by prepositions (23%), while deletions are more diverse, with a 21% article share. The frequency of article errors is likely to depend on the L1 of the speaker. Slavic languages, for instance, make considerably less use of the definite article than Esperanto, and in Mandarin Chinese the closest equivalent to a definite article is a demonstrative.

Finally, we mark punctuation errors, the vast majority being comma errors (15.1% of all error markings), with sentence splitting and other punctuation accounting for under 0.5%. Comma correction has been implemented at the clause level only[18], as group level commas (e.g. lists and appositions) do not present a problem for most language users. Because they have to take into account overall sentence structure and possible word order variation, the necessary CG rules are fairly complex, not least for clause-end commas. (r7) is such a rule, targeting the left-most adjacent dependent (&>A, &>N, &ARG>, &ADVL>) of a main-clause finite verb (&FV) – or the verb itself - with a left context of a subclause finite verb (&FS) without a left argument or anouther main-clause verb coming in between. The real rule has seven NEGATE contexts as a safety measure, only two of which are shown here, a crossing prepositional argument dependency (&P<) and an un-accounted-for left subject (&SUBJ>).

(r7) ADD (@comma-FSend) TARGET &>A OR &>N OR &ARG/ADVL> OR &FV
(*1S &FV BARRIER &<ARG/ADVL OR CLB OR &MV LINK NOT 0 @inf)
(NOT -1 &>A OR &>N OR &ARG/ADVL> OR KOMMA OR CLB-ORD OR KC OR HYFEN)
(*-1 &FS BARRIER &ARG/ADVL> OR &FV LINK NOT 0 &FV)

[18] There are no official comma rules for Esperanto, but classical literature mostly uses a central-European grammatical comma separating clauses with start, end and coordination commas. This is also the editorial strategy of the international Esperanto journal *Monato* (https://www.monato.be/konvencioj.php#inter).

(NEGATE *1 &P< BARRIER PRP)
(NEGATE *-1 &SUBJ>& BARRIER VFIN)

There are five types of clause separation commas, plus a marker for a wrong/spurious comma:

Tag	Error	%
FS-start	start of subclause	64.5
FS-end	end of subclause	4.5
FSco	subclause coordination	2.8
FMco	main clause coordination	24.1
contrast	contrasting comma	2.1
comma	unspecified	0.5
no-comma	wrong/spurious comma	1.5
		100

Table 3: Comma errors

The 2/3 dominance of the subclause start comma can be explained by the fact that subclauses are more likely to be appended than prepended. In addition it may be a factor that this type of comma is not used in English-speaking countries.

4 Pedagogical considerations

Apart from a machine-learning perspective, the EspEraro corpus has mainly a pedagogical purpose, not least allowing teachers to find and contrast typical error examples, and for students, to understand usage patterns. In this context, some error types are more important than others. The arguably most stigmatizing[19] error in Esperanto is not using the accusative *-n* correctly. It is therefore an advantage that case error categories are strongly represented in the corpus (60% of all grammatical errors). In addition, the fact that 9 subtypes are distinguished (5 @acc, 4 @nom) has explicative value, as mixing examples (e.g. object with adverbial uses) would make it more difficult for the learner to grasp the usage scenarios for the accusative ending.

Another pedagogically important topic in the corpus are confusion word pairs and false friends (e.g. *letero/litero* ['mail letter' vs. 'a-z letter'], *necesi/bezoni* [be necessary vs. need]). Here, contextualized error examples (i.e. whole sentences) not only help to perceive usage differences, but also to implicitly define the two different meanings through the example's semantic context. Because of its text book sources, the corpus has a comprehensive coverage of this type of error, with ~350 individual confusion pair lemmas.

In terms of usage patterns it should be born in mind that the corpus dowith es not *only* contain erroneous uses of a problematic lemma, but also *correct* uses. Thus, the partitive preposition *'da'* is represented with 8 *de-->da* corrections, 5 insertions and 14 *da-->de* corrections, but also with 211 ordinary, correct occurrences. The material could be used to extract a usage snapshot or to create insertion exercises, like the red 7x7 and 8x8 combination matrices below (e.g *sufiĉe/multe/... de (→ da) seĝoj/mono/...* etc.):

{sufiĉe | multe | miliono | deko | guto | tino | sako}
[enough, much, million, ten_noun, drop, vat]

de → da

{seĝoj | mono | infanoj | lingvoj | sango | ligno | ovoj}
[chairs, money, children, languages, blood, wood, eggs]

and/or, for the opposite confusion error:

{plimulto| plejmulto| specoj | malpermeso | ĉenoj | 68% | manko | loĝantaro}
[larger part, majority, types, interdiction, chains, 68%, lack, inhabitants]

da → de

{homoj | kubanoj | floroj | circumcido | bambuaroj | voĉdonoj | tempo | iom malpli ol}
[people, cubans, flowers, circumcision, bamboo bushes, votes, time, a little more than]

and, finally, for the omission error:

Ø → da

centoj (da) miloj [hundreds of millions]
pli (da) pomoj [more apples]
milionoj (da) gastoj [millions of guests]
pli (da) amikojn [more friends_accusative]

5 Outlook

With the large scope of manually revised error types contained in the current version of the corpus, and given the detailed CG analysis of error contexts, we hope EspEraro will fulfill its purpose as a teaching and development tool.

[19] The error has its own Facebook page as "most liked and hated error": https://www.facebook.com/akuzativo

However, the corpus is too selective and too small to permit true linguistic-statistical research on the real-life distribution of Esperanto errors. So future work should trade quality against quantity for this purpose, creating a large, multi-genre sister corpus of at least 10 million words, with automatic annotation only, allowing diachronic studies, genre comparison and – given the necessary meta data - an examination of the influence of an author's L1.

Acknowledgments

Lingvohelpilo-2 is a GrammarSoft ApS project and has been made possible by funding received from the Esperantic Studies Foundation (ESF). Current development has drawn on prior work carried out during the first Lingvohelpilo project[20] 2008-2010, as well as improvements made to the spellchecker and its underlying EspGram parser in the interim.

References

Reima Al-Jarf. 2010. *Spelling Error Corpora in EFL*. 7. 6-15. 10.17265/1539-8072/2010.01.002.

Þórunn Arnardóttir, Isidora Glisic, Annika Simonsen, Lilja Stefánsdóttir, and Anton Ingason. 2022. Error Corpora for Different Informant Groups: Annotating and Analyzing Texts from L2 Speakers, People with Dyslexia and Children. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pp. 245–252, New Delhi, India. ACL

Eckhard Bick. 2007. Tagging and Parsing an Artificial Language: An Annotated Web-Corpus of Esperanto, In: *Proceedings of Corpus Linguistics 2007, Birmingham, UK*. Electronically published at (http://ucrel.lancs.ac.uk/publications/CL2007/, Nov. 2007)

Eckhard Bick. 2023. VISL & CG-3: Constraint Grammar on the Move: An application-driven paradigm. In: Arvi Hurskainen, Kimmo Koskenniemi & Tommi Pirinen (eds.), *Rule-Based Language Technology.* NEALT Monograph Series vol. 2, pp. 112-140. University of Tartu. ISSN 1736-6291

M. Gamon, M. Chodorow, C. Leacock, and J. Tetreault. 2013. Using learner corpora for automatic error detection and correction. In A. Diaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam and Philadelphia: John Benjamins, pp. 127–149.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora Generation for Grammatical Error Correction. In Proceedings of NAACL 2019: Human Language Technologies, Vol. 1, pp. 3291–3301, Minneapolis, Minnesota. ACL.

Ekaterina Rakhilina, Anastasia Vyrenkova, Elmira Mustakimova, Alina Ladygina, and Ivan Smirnov. Building a learner corpus for Russian. 2016. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition* at SLTC, Umeå, 16th November 2016. http://aclweb.org/anthology/W16-65

[20] For information about this project and access to the old tool, see https://edu.visl.dk/lingvohelpilo/