A Survey of Error Annotation Schemes for Human and Machine Generated Text

Rudali Huidrom and **Anya Belz**
ADAPT Research Centre, Dublin City University
{rudali.huidrom,anya.belz}@adaptcentre.ie

Abstract

While automatically computing numerical scores remains the dominant paradigm in NLP system evaluation, error annotation and analysis is receiving increasing attention, with several error annotation schemes recently proposed for automatically generated text. However, there is little agreement about what error annotation schemes should look like, how many different types of errors should be distinguished and at what level of granularity. In this paper, our aim is to map out work on annotating errors in human and machine generated text, with a particular focus on error taxonomies. We describe our paper selection process, and survey the error annotation schemes reported in the papers, drawing out similarities and differences between them. Finally, we characterise the issues that would make it difficult to move from the current situation to a standardised error taxonomy for annotating errors in automatically generated text.

1 Introduction

Error analysis and reporting is commonly encouraged in the natural language processing (NLP) field to aid in understanding system weaknesses, most recently those that are exhibited by state-of-the-art neural systems (van Miltenburg et al., 2021), which have led to renewed calls in NLP for error analysis and building error taxonomies (Costa et al., 2015; Rivera-Trigueros, 2021).

With the advancement of neural networks and growing interest beyond pipeline-based approaches, semantic errors are increasingly observed in generation scenarios. In data-to-text generation, for example, about 40% of the E2E Generation Challenge system outputs contained erroneously omitted or added semantic content (Dušek et al., 2020). Ideally, the data-to-text systems that we develop produce outputs that convey all and only the input content (not omitting or arbitrarily adding any content) (Dušek et al., 2019; Harkous et al., 2020), so it is important to identify and understand what kinds of semantic errors occur and for what reasons, for which error annotation and subsequent analysis provides a basis. However, there is currently little agreement on how the annotation part of this should be done.

In this paper, we present a survey of different error annotation schemes, with a particular focus on error taxonomies, that have been proposed in NLP. Our paper selection process yielded a set of 22 papers reporting error type annotations and error taxonomies from the ACL Anthology. The scope of this paper is limited to error annotation schemes that include semantic errors (as well as, possibly, other types of errors, e.g. syntactic and commonsense errors).

The paper is organised as follows: Section 2 describes the paper selection and filtering process. Section 3 provides summaries of the research presented in each paper. Section 4 presents a comparative survey of the papers in terms of shared properties, Section 5 discusses our findings, and Section 6 concludes with a summary and future directions.

2 Paper Selection and Filtering

To select papers for our survey, we searched the ACL Anthology[1] with the query terms "error taxonomy" and "NLP," and "error type annotation" and "NLP" which yielded 84 results. After removing non-paper results and duplicates,[2] we were left with 27 papers. We manually examined the remaining papers keeping only those that actually reported an error taxonomy or error annotation scheme including semantic errors, which left 18 papers. We added four relevant papers from the related work

[1] https://aclanthology.org

[2] Search results included 39 author profiles, and 18 paper duplicates, where papers are repeated in two places, e.g. when the same paper is found both individually and as a part of proceedings in the search.

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 383 - 398
December 7, 2022 ©2022 Association for Computational Linguistics

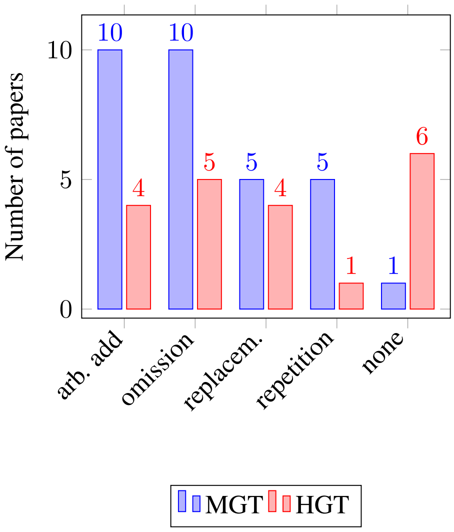


Figure 1: Number of papers reporting semantic errors in the taxonomies.

sections of three of the 18 papers. Our survey reviews the resulting 22 papers.

3 Paper Summaries

In this section, we provide high-level summaries of each of the 22 papers in our survey.

Costa et al. (2015) propose a linguistically motivated taxonomy for machine translation (MT) errors, classifying MT errors for English to European Portuguese translation. At the top level, the taxonomy is divided into five categories: orthography, lexis, grammar, semantic and discourse. It has five levels and 25 leaf nodes. The taxonomy is shown in full in Figure 3, alongside examples extracted from the paper, in Appendix A.

Extend earlier work (Specia et al., 2021), Al Sharou and Specia (2022) propose a an error annotation scheme with seven categories of critical errors, i.e. those with potential negative impact on users where the meaning of the target text deviates drastically from the source text. The work distinguishes three ways in which meaning can deviate from the source sentence: mistranslation, hallucination and deletion. A few examples extracted from the paper are in Appendix A.

Caseli and Inácio (2020) report an error analysis for neural machine translation (NMT) system outputs for Brazilian Portuguese in which errors by the NMT system are compared with those by a PBSMT system trained on the same corpus. The paper adopts the error taxonomy from Martins and Caseli (2015) which divides errors at the top level into four broad categories: syntactic errors, lexical errors, errors involving n-grams and reordering. The taxonomy has three levels and 12 leaf nodes. The taxonomy can be found in Appendix A.

Federico et al. (2014) propose a statistical framework for analysing the impact of different error types based on the results from MT evaluation metrics and human perceptions with linear mixed effects models. Experiments are carried out for English as the source and other languages that are distant from English as the target. This paper uses a set of four general error classes: (i) reordering errors (ii) lexicon errors, (iii) missing words and (iv) morphology errors.

He et al. (2021) report an error-annotated dataset called TGEA which has comprehensive annotations for texts generated with pretrained language models. It is also intended as a benchmark dataset for automatic diagnostic tasks such as error detection, error classification etc. The error taxonomy covers 25 error types in a 3-level hierarchy reflecting linguistic knowledge as shown in Figure 4 in the Appendix.

Belkebir and Habash (2021) report an automatic error type annotation system called ARETA for Modern Standard Arabic. ARETA aims to annotate and evaluate the quality of system outputs. First, it performs word alignment of the source and the target sentence. Second, the alignment is fed to the automatic error type annotation where the system tries to extract the error type. The ARETA taxonomy is based on the Arabic Learner Corpus (ALC) error tagset (Alfaifi and Atwell, 2015) with extended merge and split classes. The latter includes 29 error tags for Arabic of which 26 are used for ARETA. The ARETA taxonomy has three levels and 26 leaf nodes. The taxonomy can be found in Figure 5 in Appendix A.

Huang et al. (2020) introduce PolyTope which quantifies primary sources of errors for 10 representative models for text summarisation. While this is not an error taxonomy paper, it reports primary sources of errors with 8 'fluency' and 'accuracy' type metrics. For PolyTope, (i) Accuracy-related issues are defined as the summarisation not matching or accurately reflecting the source text, whereas (ii) Fluency-related issues are defined as problems with the linguistic qualities of the text. Level of severity is additionally marked as minor, major or critical. The paper uses the CNN/DM Dataset (in its non-anonymous version) for experiments. The taxonomy has three levels and eight leaf nodes, as

Error Types	Definition(s)
Hallucination[a] (*h*)	An example definition in the context of NLP is "generated content that is nonsensical or unfaithful to the source content." This is a widely accepted term (Ji et al., 2022) to refer to content in the output that does not have corresponding content in the input. The term comes from the field of psychology where e.g. (Blom, 2010) defines hallucination as "a percept, experienced by waking individual, in the absence of an appropriate stimulus from extracorporeal world." Some other terms used for errors very similar to hallucination are *addition, insertion, extra words, unnecessary information.*
Omission (*o*)	Used so commonly in MT that a definition is not usually given, this term refers to content in the input that should be rendered in the output not having corresponding content in the output (Weng et al., 2020) Some alternative terms in use are *deletion, absent word/n-gram, missing context/information.*
Replacement (*re*)	We use this term to refer to a range of error phenomena (given a variety of names in the literature) where some content in the output is clearly intended to convey some part of the input, but does so incorrectly (Subramaniam et al., 2009; Gouws et al., 2011; Han and Baldwin, 2011; Al Sharou and Specia, 2022). We can also look at replacement errors from the perspective of a combination of omission and addition, in the special case where what is added is the incorrect version of what is omitted. Some alternative terms also in use are *substitution, mistranslation, transposition*.
Repetition (*r*)	An example definition is "occurrence of the same words several times or syntactically similar units unintentionally or on purpose" (Al Sharou et al., 2021). Some alternative terms in use are *duplication, redundancy*.

Figure 2: Definitions of high-level semantic error types found in the literature (Col 1: semantic error types; Col 2: definitions).

[a] We generally prefer the term 'arbitrary content addition' but use the original term used in the literature to avoid confusion.

showin in Figure 6.

Di et al. (2019) report a detailed analysis of errors from four morphological inflection systems for Tibetan, using datasets developed by Cotterell et al. (2018), and the error taxonomy reported by Gorman et al. (2019) for target errors and prediction errors with a more detailed analysis on (i) errors due to words that violate lexicographic or morphophonetic constraints of the language, and (ii) allomorphy errors. This latter taxonomy has three levels and three leaf nodes, and can be found in Appendix A.

Mahmud et al. (2021) report a qualitative investigation of errors made by neural models fro which they create a taxonomy which consists of seven top level categories each with multiple lower level subcategories as shown in Figure 7 in the Appendix. Altogether there are three levels and 31 leaf nodes.

Costa et al. (2012) report a corpus of about 6,000 questions manually translated into Portuguese. They provide translation guidelines which discuss two types of problems: semantic level issues and structure level issues. In addition, they report an error taxonomy with four broad error categories and carry out an error analysis. The taxonomy has three levels and nine leaf nodes, and can be found in Appendix A.

Macklovitch (1991) introduces an error taxonomy to help with post-editing operations. The error taxonomy distinguishes three broad categories at the top level: (i) Morphology, (ii) Source language analysis and (iii) Transfer and Generation. Altogether there are three levels and 19 leaf nodes. Figure 9 shows the taxonomy extracted from the paper.

Lin et al. (2022) address automatic translation error correction (TEC) where the goal is to produce an improved translation by correcting errors found in a translation. The paper proposes a pretraining approach for TEC and also introduces a human-in-the-loop user study where it was found that higher quality translations were achieved when corrections are assisted by the TEC model. The taxonomy used has three levels and five leaf nodes. It can be found in Appendix B.

van der Goot et al. (2018) describe an error taxonomy for lexical normalisation replacements. The work makes a clear distinction between intentional and unintentional anomalies, and the taxonomy has four levels and 14 leaf nodes.

Ng et al. (2014) provide an error annotation scheme for grammar error types. The paper's goal is to evaluate algorithms and systems for automatically detecting and correcting grammatical errors present in English essays written by second language learners of English. The error annotation scheme has a set of 28 categories of grammatical error corrections as a part of the CoNLL-2014 shared task. The authors report that it is often acceptable to have multiple and different corrections in grammatical error correction. The dataset used for training is the NUCLE corpus, the NUS corpus for Learner English (Dahlmeier et al., 2013), and the test data is collected as written essays from 25

NUS students who are non-native speakers of English where each student was asked to write two essays. Figure 12 in Appendix B shows the error categories from the paper.

Dickinson and Herring (2008) report a computer-aided language learning (ICALL) system for beginner-level learners of Russian. The goal of the system is to provide exercises supporting basic grammar learning with contextualisation for morphological errors. Considering the nature of the learner's language, an error taxonomy with four broad categories for Russian verbal morphology is reported. It has three levels and nine leaf nodes. More details of the taxonomy can be found in Appendix B.

Dickinson (2010) reports work on generating linguistically informed morphological errors for Russian. An error taxonomy is reported that helps in the error generation process. It has four levels and 10 leaf nodes, and can be found at Figure 13 in Appendix B.

Nagata et al. (2018) explore the influence of spelling errors on lexical variation measures like Type-Token Ratio (TTR) and Yule's K for learner English. The error annotation scheme reported presents two ways of spelling error correction: (i) it identifies 13 errors in the corpora. (ii) it classifies them into three groups: corrected, not corrected and not counted. The scheme has 13 error types. Figure 14 in Appendix B shows the list of errors from the paper.

Gayo et al. (2016) propose the COPLE2 corpus which is a new learner corpus for Portuguese. Three different linguistic error types are defined for error tagging: orthographic, grammatical and lexical. The first error type covers spelling errors, with errors here restricted to word form and punctuation marks. The second error type is for when the student has produced an ungrammatical utterance, thus going beyond individual words and considering syntactic structures. The third error type covers lexical/semantic errors. The work is mainly concerned with errors that affects meaning. These error types help in visualising the same text progressing through corrections at different stages, from the version closest to original (orthographic corrections) to the most modified one (orthographical, grammatical and lexical corrections). Note that the sub-error types provided for this paper are unclear, and are not counted in tables below.

Barbagli et al. (2016) present a collection of essays called CItA corpus written by Italian L1 learners (Corpus Italiano di Apprendenti L1) from the first and second years of lower secondary school. In addition, they report a three-level error annotation scheme for errors made by L1 Italian learners: (i) macro-class of error (grammatical, othrographic and lexical); (ii) class of error (verb, prepositions, monosyllables); and (iii) type of modification (misuse of verb with respect to verbal tense). There are therefore four levels in the underlying taxonomy, and a total of 21 leaf nodes. Figure 15 in Appendix B shows the taxonomy from the paper.

Himoro and Pareja-Lora (2020) propose a spelling error taxonomy for Zamboanga Chabacano (ZC) formalised as an ontology and an adaptive spell checking approach using character-based statistical MT. First, an iterative process is applied to samples of the CWZCC corpus for categorising different spelling errors. Second, the errors are classified to create an error taxonomy. It is observed that spelling errors get more complex as one goes deeper down the tree. The taxonomy has eight levels and 14 leaf nodes. and is shown in Figure 16 in Appendix B.

Caines et al. (2020) introduce a corpus of one-to-one online chatroom conversations from lessons between teachers and learners of English which is known as the Teacher-Student Chatroom Corpus (TSCC). A set of 24 error types is determined on the basis of the grammatical error correction of texts in the corpus. The set is shown in Figure 17 in Appendix B.

Korre et al. (2021) introduce ERRANT which is a toolkit that annotates texts and offers error typing with detailed feedback for L2 learners of Greek. Annotation is based on a rule-based error type framework that distinguishes (i) error description (Unnecessary, Replacement and Missing), and (ii) error type. The latter disinguishes 16 error types.

4 Properties of Error Annotation Schemes

In order to be able to compare different error annotation schemes and draw conclusions about their similarities and differences, we labeled each scheme in terms of (i) whether it was designed for machine or human generated text; (ii) whether it contained error types related to semantic accuracy, fluency or both; (iii) NLP system task; (iv) purpose of the annotation that was carried out; and (v) details of how many error labels and how many hierarchical levels there are in the scheme. We describe the first two in Section 4.1 below, the third and fourth in Section 4.3, and the last in Section 4.2.

The results of labelling the 22 annotation schemes with these labels are presented in Section 4.4.

4.1 Text type and error type

We categorise each paper in terms of the (i) *text type*, and (2) *error type* addressed. Regarding the former, we group the error annotation schemes in our survey into those developed for machine-generated text (MGT) and those developed for human-generated text (HGT), according to the following definitions:

- **Machine generated texts (MGT)** i.e. synthetic texts generated by a system or a model based on pre-defined rules or algorithms, including MT, text summarisation, story generation etc. Errors like mistranslation, omission or arbitrary content addition, etc. are observed frequently in annotation schemes for this text type. Figure 7 in the Appendix provides examples of typical errors.
- **Human generated texts (HGT)** include reference texts for evaluating systems, and training corpora for various downstream NLP tasks. The nature of the text depends on its intended purpose and oftentimes, they are used for evaluation of the models we build. Compared to MGT, HGTs are less prone to semantic errors. However, it cannot be generalised that all human generated texts are of good quality, and semantic errors do occur. Figure 16 in the Appendix provides examples of errors in human-generated texts.

We further categorise error annotation schemes in terms of the broad error type(s) addressed. Here we use 'accuracy' and 'fluency' as shorthand for content type errors as per Figure 2, and non-content type errors, respectively. These two terms are used frequently in the MT literature, e.g. in the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014) where they are defined as follows:

- **Accuracy**: Errors where the target sentence does not correspond to the source text due to omission, distortion or addition to the text. Error types include mistranslation, over-translation, under-translation, untranslated, omission, and addition.
- **Fluency**: Errors related to grammar and style. Examples include errors relating to spelling, punctuation, grammatical rules, inconsistent style, unidiomatic style etc.

4.2 Structure of annotation scheme

We also categorised error annotation schemes in terms of two structural properties:

1. The **number of different error types** included in an annotation scheme, for which we use a standardised definition as the number of nodes in the tree including the root;
2. The **depth of the hierarchical structure** underlying the scheme. If there is no underlying hierarchical structure, then depth=1. Depth = levels - 1, where levels are the number of nodes in the longest path from root to the leaf nodes.

4.3 NLP task and annotation scheme purpose

We distinguish the following **NLP System Tasks**, abbreviated as indicated in square brackets in tables below: Machine Translation [MT], Text Summarisation [TS], Textual Summarisation of source code [TS(SC)], Type-level Universal Morphological Reinflection Task [MI], Automatic Translation Error Correction [EC(T)], Text Normalisation [TN], Grammar Error Correction [EC(G)], Morphological Error Detection and Classification [MDC], Error Generation [EG], None (Corpus Linguistics) [N(CL)], Dialogue [D], Error Type Classification [ETC] and Spelling Error Correction [EC(S)].

NLP System Task also includes the following automatic forms of error annotation: Automatic Error Annotation for Dataset Creation [EA(D)], Automatic Error Annotation of System outputs for evaluation [EA(S)], Automatic Corpus Error annotation/analysis [CE]. The NLP System Task is defined for what the error annotation scheme is used for in its respective papers.

Inspired by Machine Translation (MT) research which takes a very structured approach to error analysis (Stymne and Ahrenberg, 2012; Koponen, 2010), error classification (Vilar et al., 2006; Popović and Burchardt, 2011; Popović, 2021), and building error taxonomies (Costa et al., 2015; Al Sharou and Specia, 2022), we also categorise

Sl.no	Work	Type	NLP Task	Accuracy	Fluency	# Error types	Depth	Purpose
1.	Costa et al. (2015)	MGT	MT	✓	✓	36	4	EAn+EA+EC
2.	Al Sharou and Specia (2022)	MGT	MT	✓		8	1	EAn+E(S)
3.	Caseli and Inácio (2020)	MGT	MT	✓	✓	17	2	EAn+EA+E(S)
4.	Federico et al. (2014)	MGT	MT	✓	✓	5	1	EAn+EA
5.	He et al. (2021)	MGT	EA(D)	✓		32	2	EAn+EA+E(S)
6.	Belkebir and Habash (2021)	MGT	EA(S)		✓	34	2	EAn+EA
7.	Huang et al. (2020)	MGT	TS	✓	✓	11	2	EAn+E(S)
8.	Di et al. (2019)	MGT	MI		✓	5	2	EAn+EA+E(S)
9.	Mahmud et al. (2021)	MGT	TS(SC)	✓		39	2	EAn+E(S)
10.	Costa et al. (2012)	MGT	CE	✓	✓	11	2	EAn+EA+E(C)
11.	Macklovitch (1991)	MGT	MT	✓	✓	23	2	EAn+E(S)
12.	Lin et al. (2022)	HGT	EC(T)		✓	7	2	EAn+E(C)+E(S)
13.	van der Goot et al. (2018)	HGT	TN		✓	20	3	EAn+E(S)
14.	Ng et al. (2014)	HGT	EC(G)		✓	29	1	EAn+E(S)
15.	Dickinson and Herring (2008)	HGT	MDC		✓	11	2	ED+EA
16.	Dickinson (2010)	HGT	EG		✓	14	3	EAn+E(C)
17.	Nagata et al. (2018)	HGT	EC(S)	✓	✓	14	1	EAn+E(C)+EC
18.	Gayo et al. (2016)	HGT	CE	✓	✓	4	1	EAn+E(C)
19.	Barbagli et al. (2016)	HGT	N(CL)	✓	✓	35	3	EAn+E(C)
20.	Himoro and Pareja-Lora (2020)	HGT	EC(S)		✓	39	7	EAn+E(C)+EC
21.	Caines et al. (2020)	HGT	D		✓	25	1	EAn+E(C)
22.	Korre et al. (2021)	HGT	ETC		✓	17	1	EAn+E(C)+EC

Table 1: Overview table of properties of the error annotations schemes surveyed (for explanation of abbreviations see Table 3 and in text).

our error annotation schemes in terms of the **Purpose for which an error annotation scheme was created** as follows: Error Classification is EC, Error Annotation is EAn, Evaluation for systems is E(S), Evaluation for corpus errors is E(C), Error Detection is ED and Error Analysis is EA. The purpose is defined for what the error annotation scheme that was created is used as in its respective papers.

4.4 Labelled annotation schemes

Table 1 shows each of the 22 surveyed papers alongside their individual labels. Columns 3 and 4 indicate text type and NLP Task, Columns 5 and 6 whether Accuracy or Fluency is addressed, and the last three columns show number of different Error Types, Depth, and Purpose for which the scheme was created, respectively, all as defined in the preceding section.

As can be seen, there is an even distribution of papers over text type addressed (HGT vs. MGT). Moreover, none of the 11 papers addressing HGT address only accuracy errors, most address only fluency (8 out of 11), and just three address both accuracy and fluency errors. For the 11 MGT papers, we have a fair mix of different types of errors i.e., three address only accuracy errors, two only fluency errors, and six address both. Determining the number of error types and the depth of the hierarchy (if any) has been a challenge due to lack of clarity within the papers. For example, Gayo et al. (2016) do not mention the error sub-types in the taxonomy clearly which makes counting them difficult. This means we have provided an estimate in some cases.

All 22 papers have a combination of purposes for which the scheme was created (last column).

5 Discussion

5.1 Trends Observed

Table 2 presents the overall trend in different types of semantic errors included in error annotation schemes over the years in our surveyed papers. We mark as 1 if we encounter any one of the semantic error types from Figure 2 in a paper (each paper can have more than one semantic error type). For example, we have a count of 4 for arbitrary content addition errors from 2020 which means four papers

Semantic Error	1991	2008	2010	2012	2014	2015	2016	2018	2019	2020	2021	2022	TOTAL
Arbitrary content addition (*h*)	1			1	1	1	1	1		4	4	1	15
Omission (*o*)	1			1	1	1		1		4	4	1	14
Replacement (*re*)	1				1					4	2	2	10
Repetition (*r*)						1	1			2	2	1	7
TOTAL	**3**			**2**	**3**	**3**	**2**	**2**		**14**	**12**	**5**	**46**

Table 2: Number of taxonomies that incorporated each of the four high-level semantic error types from Figure 2, shown per publication year in our set of 22 papers.

Purpose	(Paper)[Sem. Error]	MGT	HGT
Error Analysis & Error Annotation (EA+EAn)	(4,6)[*h,o,re*]	✓	
Error Detection & Error Analysis (ED+EA)	(15)[*n*]		✓
Evaluation of systems & Error Annotation (E(S)+EAn)	(2,7,9,11,13)[*h,o*],(2,11)[*re*], (7,9)[*r*],(14)[*n*]	✓	✓
Error Annotation & Error Analysis & Error Classification (EAn+EA+EC)	(1)[*h,o,r*]	✓	
Evaluation of corpus errors & Error Annotation (E(C)+EAn)	(19,21)[*o*],(21)[*h,re*], (16,18)[*n*]		✓
Error Annotation & Error Analysis & Evaluation of systems (EAn+EA+E(S))	(3,5)[*h,o,r*],(3)[*re*],(8)[*n*]	✓	
Error Annotation & Error Analysis & Evaluation of corpus errors (EAn+EA+E(C))	(10)[*h,o*]	✓	
Error Annotation & Evaluation of corpus errors & Evaluation of systems (EAn+E(C)+E(S))	(12)[*re*]		✓
Error Annotation & Evaluation of corpus errors & Error Classification (EAn+E(C)+EC)	(20,22)[*h,o,re*], (17)[*n*]		✓

Table 3: For each (combination of) purpose(s) in the 22 surveyed papers, the taxonomies to which it applies (round brackets), and the semantic error types covered by each of those taxonomies [square brackets]. We also show text type to which each (combination of) purpose(s) applies.

address such errors in the year 2020. The paper IDs are taken from Table 1.

Four out of the five papers in our survey published more than ten years ago (2012 and earlier) are categorised as HGT (except the paper by Macklovitch (1991) which is categorised as MGT), and do not report any semantic errors in their error annotation scheme. In addition, another paper, by Di et al. (2019), grouped under MGT, also does not report any semantic errors. We observe a total of 46 semantic error types reported in the papers from our survey.

Table 3 shows in the first column, all the combinations of purposes for which an error annotation scheme was created that we encountered in our 22 surveyed papers. The second column shows paper number (e.g. "(15)"), type of semantic errors addressed in each paper (e.g. "[*h, o, r*]") or none ("[*n*]"). The last two columns show text type (HGT vs. MGT).

We observe that papers where text type is MGT typically address one or more semantic errors (10 out of 11 papers), except for the paper by Di et al. (2019) whose purpose is error analysis and evaluation of systems. Half of the papers labelled HGT do not address any semantic errors. The other half of the papers with error annotation and evaluation of corpus or error classification as purpose in HGT addresses semantic errors. The statistics of how many papers address each of the high-level semantic error types in the MGT and HGT groups can easily be seen in Figure 1.

Table 2 shows the high-level semantic error types reported in each year in our survey. It is interesting to observe that addressing semantic errors has become increasingly frequent in very recent years.[3] One reason is likely to be the shift from controlled pipeline approaches to end-to-end neural approaches for many NLP tasks.

[3] Note that we performed the ACL Anthology search in August, so we may be missing some papers from 2022.

5.2 Observations

In this section, we discuss the issues we observed in labelling the error annotation schemes and taxonomies in our survey. We summarise these observations from the perspective of semantic errors as follows:

1. **Lack of standardisation across schemes** (e.g., definition, examples) is observed which hampers deriving a standardised framework for semantic errors. We found that 11 out of 22 papers (50%) mention only the name of an error type or its sub-type without defining them at all. It is highly observed in the HGT group, with eight out of 22 papers. In some cases it is difficult to categorise schemes/taxonomies in terms of the *high-level error type(s)* from Figure 2.
2. **Differing and/or incompatible error names and definitions:** In our survey, we encountered only two papers (Dickinson and Herring, 2008; Dickinson, 2010) with mutually compatible error type definitions and these are by the same first author. For the remaining papers, either the error definitions means the same but the error term is different, or vice versa.
3. **Borderline error types** that cannot clearly be assigned either to semantic accuracy or to fluency. Categorising the error annotation schemes for which this is the case in the survey as accuracy and/or fluency errors is sometimes difficult due to the (lack of) provided definitions, examples, etc. We found 12 out of 22 papers (which is more than 50%) to be difficult to categorise which corresponds to three out of 11 papers for the MGT group, and 9 out of 11 papers for HGT. This difficulty implies an unclear boundary between accuracy and fluency types of errors.

Further interesting observations can be drawn from Table 1 and Table 3 concerning the relationship between semantic errors on the one hand, and text types (MGT, HGT) and broad errors types (accuracy and fluency), on the other. Nine out of 11 papers in the MGT group address either accuracy error types only (3/9), or accuracy error types together with fluency error types (6/9). Arbitrary content addition (currently more commonly known as hallucination) and omission are the most common semantic errors reported and we see them in all nine papers under the MGT group. Repetition (in combination with other semantic errors) is the next frequently reported semantic error and we see it in five out of nine papers under the MGT group. Meanwhile, three papers in the HGT group address both accuracy and fluency error types with only one paper out of the three addressing omission.

Part of the motivation for conducting this survey was to use it as a starting point in creating our own semantic error taxonomy for annotating errors in output text in data-to-text generation. Our original aim was to base our taxonomy on common error types found in the literature, but, as we have seen in this paper, there is little agreement between existing error taxonomies beyond a distinction at the highest level between accuracy and fluency type errors, and accuracy further dividing into (a) arbitrarily content addition (currently more commonly known as hallucination), (b) omission, (c) replacement, and (d) repetition. Our next step will be to take this as a starting point and add lower taxonomy levels as required for data-to-text generation, while trying to incorporate as much common ground from the literature as possible. Another consideration will be that we wish to use the resulting error taxonomy both in performing manual error analysis, and for providing the categories in automatic error detection.

6 Conclusion

We conducted a structured survey of work on error type annotation schemes (with a focus on error taxonomies), as reported in papers from the ACL Anthology. We observed a number of issues while analysing the papers in our survey which we characterised in terms of (1) lack of standardisation, (2) differing/incompatible error names and definitions across different papers, and (3) borderline error types which resist being classified as either fluency or accuracy related. We found that the latter is mostly observed in error annotation for human-generated text rather than machine-generated. We conducted our study from the perspective of semantic error annotation, as we plan to build on it in future work on developing an error taxonomy of semantic error types for data-to-text generation.

References

Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. Towards a better understanding of noise in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62.

Khetam Al Sharou and Lucia Specia. 2022. A taxonomy and study of critical errors in machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–179.

AYG Alfaifi and ES Atwell. 2015. Computer-aided error annotation: a new tool for annotating arabic error.

Alessia Barbagli, Pietro Lucisano, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2016. Cita: An l1 italian learners corpus to study the development of writing competence. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 88–95.

Riadh Belkebir and Nizar Habash. 2021. Automatic error type annotation for arabic. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 596–606.

Jan Dirk Blom. 2010. *A dictionary of hallucinations*. Springer.

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. The teacher-student chatroom corpus. In *Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning*, pages 10–20.

Helena Caseli and Marcio Inácio. 2020. Nmt and pbsmt error analyses in english to brazilian portuguese automatic translations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3623–3629.

Ângela Costa, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29(2):127–161.

Ângela Costa, Tiago Luís, Joana Ribeiro, Ana Cristina Mendes, and Luísa Coheur. 2012. An english-portuguese parallel corpus of questions: translation guidelines and application in smt. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2172–2176.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 22–31.

Qianji Di, Ekaterina Vylomova, and Timothy Baldwin. 2019. Modelling tibetan verbal morphology. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 35–40.

Markus Dickinson. 2010. Generating learner-like morphological errors in russian. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 259–267.

Markus Dickinson and Joshua Herring. 2008. Developing online icall resources for russian. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.

Ondřej Dušek, David M. Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426, Tokyo, Japan. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language*, 59:123–156.

Marcello Federico, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2014. Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1643–1653.

Iria Del Río Gayo, Sandra Antunes, Amália Mendes, and Maarten Janssen. 2016. Towards error annotation in a learner corpus of portuguese. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 8–17.

Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. Weird inflects but OK: Making sense of morphological generation errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151, Hong Kong, China. Association for Computational Linguistics.

Stephan Gouws, Dirk Hovy, and Donald Metzler. 2011. Unsupervised mining of lexical variants from noisy text. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 82–90.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th annual meeting of the* *association for computational linguistics: Human language technologies*, pages 368–378.

Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2410–2424, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jie He, Bo Peng, Yi Liao, Qun Liu, and Deyi Xiong. 2021. Tgea: An error-annotated dataset and benchmark tasks for textgeneration from pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6012–6025.

Marcelo Yuji Himoro and Antonio Pareja-Lora. 2020. Towards a spell checker for Zamboanga Chavacano orthography. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2685–2697, Marseille, France. European Language Resources Association.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? *arXiv preprint arXiv:2010.04529*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*.

Maarit Koponen. 2010. Assessing machine translation quality with error analysis. In *Electronic proceeding of the KaTu symposium on translation and interpreting studies*.

Katerina Korre, Marita Chatzipanagiotou, and John Pavlopoulos. 2021. Elerrant: Automatic grammatical error type classification for greek. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 708–717.

Jessy Lin, Geza Kovacs, Aditya Shastry, Joern Wuebker, and John DeNero. 2022. Automatic correction of human translations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–507.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*, (12):455–463.

Elliott Macklovitch. 1991. Evaluating commercial mt systems. In *Evaluators' Forum on MT systems, organized by ISSCO at Ste. Croix, Switzerland.*

Junayed Mahmud, Fahim Faisal, Raihan Islam Arnob, Antonios Anastasopoulos, and Kevin Moran. 2021. Code to comment translation: A comparative study on model effectiveness & errors. In *Proceedings of the 1st Workshop on Natural Language Processing for Programming (NLP4Prog 2021)*, pages 1–16.

Débora Beatriz de Jesus Martins and Helena de Medeiros Caseli. 2015. Automatic machine translation error identification. *Machine Translation*, 29(1):1–24.

Ryo Nagata, Taisei Sato, and Hiroya Takamura. 2018. Exploring the influence of spelling errors on lexical variation measures. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2391–2398.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Maja Popović. 2021. On nature and causes of observed mt errors. In *Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track)*, pages 163–175.

Maja Popović and Aljoscha Burchardt. 2011. From human to automatic error classification for machine translation output. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*.

Irene Rivera-Trigueros. 2021. Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, pages 1–27.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André FT Martins. 2021. Findings of the wmt 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725.

Sara Stymne and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1785–1790.

L Venkata Subramaniam, Shourya Roy, Tanveer A Faruquie, and Sumit Negi. 2009. A survey of types of text noise and techniques to handle noisy text. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, pages 115–122.

Rob van der Goot, Rik van Noord, and Gertjan van Noord. 2018. A taxonomy for in-depth evaluation of normalization for user generated content. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

David Vilar, Jia Xu, Luis Fernando d'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the fifth international conference on language resources and evaluation (LREC'06)*.

Rongxiang Weng, Heng Yu, Xiangpeng Wei, and Weihua Luo. 2020. Towards enhancing faithfulness for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2675–2684.

A Appendix: Error Annotation Schemes for Machine Generated Text (MGT)

In this section, we present entire taxonomies and examples of individual errors for error annotation schemes developed for HGT and extracted from the 22 surveyed papers briefly summarised in Section 3.

Costa et al. (2015) present the taxonomy found in Figure 3. Below are some error examples extracted from the paper:

Example 1: Spelling error in Orthography level

Example: Spelling error

EN: Basilica of the Martyrs

EP: Basílica dos *Mátires

Correct translation: Basílica dos Mártires

Example 2: Omission error (content word) in Lexis level

Example: Omission error (content word)

EN: In his inaugural address, Barack Obama

EP: No seu * inaugural, Barack Obama

Correct translation: No seu discurso inaugural, Barack Obama

Example 3: Addition error (content word) in Lexis level

Example: Addition error (content word)

EN: This time I'm not going to miss

EP: Desta vez *correr não vou perder

Correct translation: Desta vez não vou perder

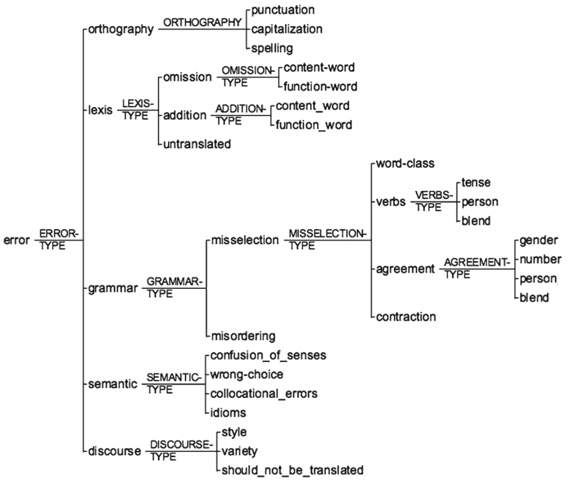


Figure 3: Figure of taxonomy extracted from (Costa et al., 2015).

Al Sharou and Specia (2022) define seven main categories as mentioned in Section 3. Here are some error examples extracted from the paper:

Example 1: Deviation in toxicity (TOX)

ST: Your killing the fucking planet.

MT-ed text: May the damn planet kill you.

Translation into Arabic by Systran

Example 2: Deviation in health/safety risks (SAF)

ST: I Know two teenagers that suffer from gerd it is a big problem for these people!

MT-ed text: I Know two teenagers that suffer from root disease it is a big problem for these people!

Translation into Chinese by GT.

Example 3: Deviation in named entities (NAM)

ST: Your fucking ass doesn't know shit about it AT ALL.Rocky.

MT-ed text: Your fucking ass doesn't know shit about it AT ALL.rock.

Translation into Italian by Bing

Caseli and Inácio (2020) presents a taxonomy found below:

1. Syntactic errors
   - Number agreement
   - Gender agreement
   - Verb inflection
   - PoS

   \item Lexical errors
   - Extra word
   - Absent word
   - Not translated word
   - Incorrectly translated word
2. N-gram
   - Absent n-gram
   - Not translated n-gram
   - Incorrectly translated n-gram
3. Reordering
   - Order

He et al. (2021) present a taxonomy found in Figure 4.

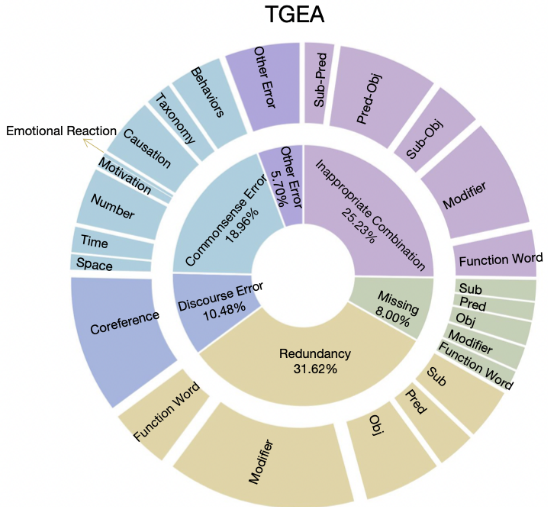
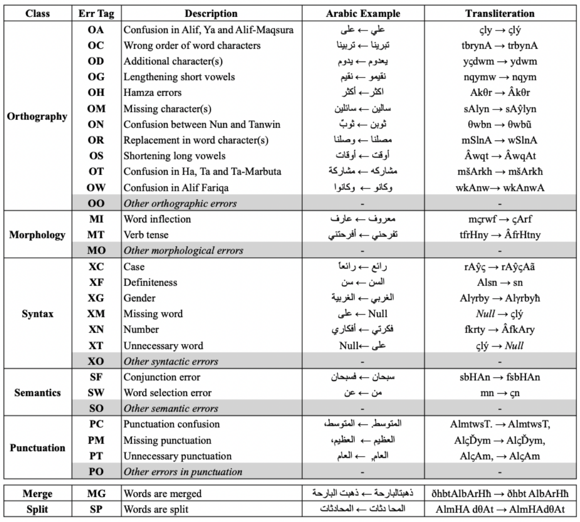


Figure 4: Figure extracted from (He et al., 2021) of level-1 and level-2 error types in TGEA which is an error annotated dataset.

Belkebir and Habash (2021) present a taxonomy found in Table Figure 5.

Huang et al. (2020) show PolyTope with each error types on a three-coordinates for syntactic and semantic roles which is found in Figure 6 and some examples extracted from the paper are found below.



Class	Err Tag	Description	Arabic Example	Transliteration
Orthography	OA	Confusion in Alif, Ya and Alif-Maqsura	علي ← على	çly → çlý
	OC	Wrong order of word characters	تبرينا ← تربينا	tbrynA → trbynA
	OD	Additional character(s)	يعدوم ← يدوم	yçdwm → ydwm
	OG	Lengthening short vowels	نقيمو ← نقيم	nqymw → nqym
	OH	Hamza errors	اكثر ← أكثر	Akθr → Âkθr
	OM	Missing character(s)	سالين ← سائلين	sAlyn → sAŷlyn
	ON	Confusion between Nun and Tanwin	ثوبن ← ثوبٌ	θwbn → θwbũ
	OR	Replacement in word character(s)	مصلنا ← وصلنا	mSlnA → wSlnA
	OS	Shortening long vowels	أوقت ← أوقات	Âwqt → ÂwqAt
	OT	Confusion in Ha, Ta and Ta-Marbuta	مشاركه ← مشاركة	mšArkh → mšArkħ
	OW	Confusion in Alif Fariqa	وكانو ← وكانوا	wkAnw→ wkAnwA
	OO	*Other orthographic errors*	-	-
Morphology	MI	Word inflection	معروف ← عارف	mçrwf → çArf
	MT	Verb tense	تفرحني ← أفرحتني	tfrHny → ÂfrHtny
	MO	*Other morphological errors*	-	-
Syntax	XC	Case	رائع ← رائعاً	rAŷç → rAŷçAã
	XF	Definiteness	السن ← سن	Alsn → sn
	XG	Gender	الغربي ← الغربية	Alγrby → Alγrbyħ
	XM	Missing word	Null ← على	*Null* → çlý
	XN	Number	فكرتي ← أفكاري	fkrty → ÂfkAry
	XT	Unnecessary word	على ←Null	çlý → *Null*
	XO	*Other syntactic errors*	-	-
Semantics	SF	Conjunction error	سبحان ← فسبحان	sbHAn → fsbHAn
	SW	Word selection error	من ← عن	mn → çn
	SO	*Other semantic errors*	-	-
Punctuation	PC	Punctuation confusion	المتوسط. ← المتوسط،	AlmtwsT. → AlmtwsT,
	PM	Missing punctuation	العظيم ← العظيم،	AlçĎym → AlçĎym,
	PT	Unnecessary punctuation	العام، ← العام	AlçAm, → AlçAm
	PO	*Other errors in punctuation*	-	-
Merge	MG	Words are merged	ذهبتالبارحة ← ذهبت البارحة	ðhbtAlbArHħ → ðhbt AlbArHħ
Split	SP	Words are split	المحا دثات ← المحادثات	AlmHA dθAt → AlmHAdθAt

Figure 5: Figure of ARETA which is an error annotation system extracted from (Belkebir and Habash, 2021).

Example 1: Inaccuracy Intrinsic

> "Pittsburgh Union Station is 10 kilometers from Exhibition Center and 3 kilometers from the University of Pittsburgh" in the source but "Pittsburgh Union Station is 3 kilometers from Exhibition Center" in the output.

Example 2: Inaccuracy Extrinsic

> it is described as "Pittsburgh Union Station, also known as Pittsburgh South Station" in the output but "Pittsburgh South Station" is neither mentioned in the source text nor exists in the real world.

Example 3: Positive-Negative Aspect

> "push a button" summarized as "don't push a button", "non-slip" summarized as "slip". This category applies only to actions and modifiers and refers to omitted or added negative particles.

Di et al. (2019) presents an error taxonomy found below.

1. Target errors: Target word errors are 'due to errors in the Wiktionary source data and incorrect extraction of paradigm tables.'
2. Prediction errors

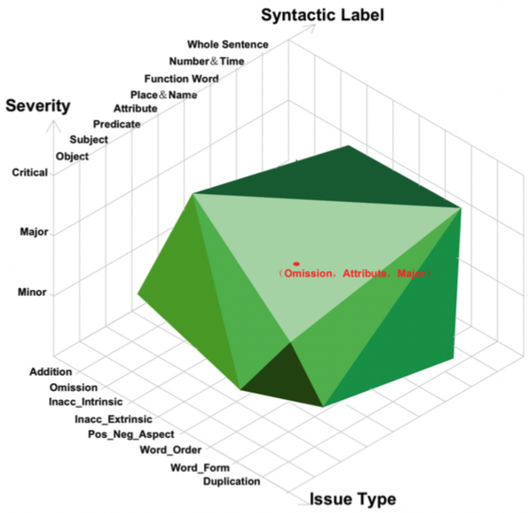
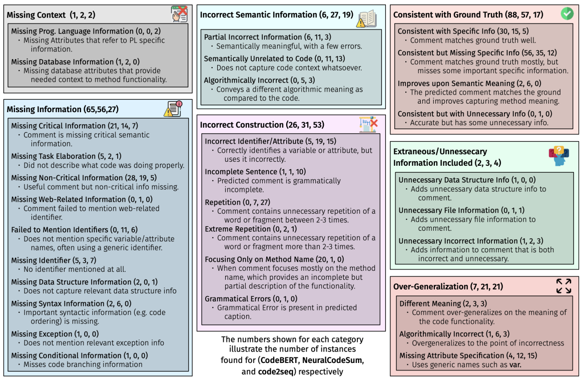


Figure 6: Figure of PolyTope with each error types on a three-coordinates for syntactic and semantic roles extracted from (Huang et al., 2020).

- nonce-word errors: Nonce word errors are 'due to illegal words, i.e. situations when the string generated by a system does not exist in Tibetan.'
- allomorphy errors: Allomorphy errors are considered for verb inflection and diacritics for Tibetan language.

Mahmud et al. (2021) present a taxonomy found in Figure 7.



Missing Context (1, 2, 2)

Missing Prog. Language Information (0, 0, 2)
- Missing Attributes that refer to PL specific information.

Missing Database Information (1, 2, 0)
- Missing database attributes that provide needed context to method functionality.

Missing Information (65,56,27)

Missing Critical Information (21, 14, 7)
- Comment is missing critical semantic information.

Missing Task Elaboration (5, 2, 1)
- Did not describe what code was doing properly.

Missing Non-Critical Information (28, 19, 5)
- Useful comment but non-critical info missing.

Missing Web-Related Information (0, 1, 0)
- Comment failed to mention web-related identifier.

Failed to Mention Identifiers (0, 11, 6)
- Does not mention specific variable/attribute names, often using a generic identifier.

Missing Identifier (5, 3, 7)
- No identifier mentioned at all.

Missing Data Structure Information (2, 0, 1)
- Does not capture relevant data structure info

Missing Syntax Information (2, 6, 0)
- Important syntactic information (e.g. code ordering) is missing.

Missing Exception (1, 0, 0)
- Does not mention relevant exception info

Missing Conditional Information (1, 0, 0)
- Misses code branching information

Incorrect Semantic Information (6, 27, 19)

Partial Incorrect Information (6, 11, 3)
- Semantically meaningful, with a few errors.

Semantically Unrelated to Code (0, 11, 13)
- Does not capture code context whatsoever.

Algorithmically Incorrect (0, 5, 3)
- Conveys a different algorithmic meaning as compared to the code.

Incorrect Construction (26, 31, 53)

Incorrect Identifier/Attribute (5, 19, 15)
- Correctly identifies a variable or attribute, but uses it incorrectly.

Incomplete Sentence (1, 1, 10)
- Predicted comment is grammatically incomplete.

Repetition (0, 7, 27)
- Comment contains unnecessary repetition of a word or fragment between 2-3 times.

Extreme Repetition (0, 2, 1)
- Comment contains unnecessary repetition of a word or fragment more than 2-3 times.

Focusing Only on Method Name (20, 1, 0)
- When comment focuses mostly on the method name, which provides an incomplete but partial description of the functionality.

Grammatical Errors (0, 1, 0)
- Grammatical Error is present in predicted caption.

The numbers shown for each category illustrate the number of instances found for (**CodeBERT**, **NeuralCodeSum**, and **code2seq**) respectively

Consistent with Ground Truth (88, 57, 17)

Consistent with Specific Info (30, 15, 5)
- Comment matches ground truth well.

Consistent but Missing Specific Info (56, 35, 12)
- Comment matches ground truth mostly, but misses some important specific information.

Improves upon Semantic Meaning (2, 6, 0)
- The predicted comment matches the ground and improves capturing method meaning.

Consistent but with Unnecessary Info (0, 1, 0)
- Accurate but has some unnecessary info.

Extraneous/Unnessecary Information Included (2, 3, 4)

Unnecessary Data Structure Info (1, 0, 0)
- Adds unnecessary data structure info to comment.

Unnecessary File Information (0, 1, 1)
- Adds unnecessary file information to comment.

Unnecessary Incorrect Information (1, 2, 3)
- Adds information to comment that is both incorrect and unnecessary.

Over-Generalization (7, 21, 21)

Different Meaning (2, 3, 3)
- Comment over-generalizes on the meaning of the code functionality.

Algorithmically Incorrect (1, 6, 3)
- Overgeneralizes to the point of incorrectness

Missing Attribute Specification (4, 12, 15)
- Uses generic names such as **var**.

Figure 7: Figure of taxonomy of errors between the generated summaries and ground truth extracted from (Mahmud et al., 2021).

Costa et al. (2012) present a taxonomy and examples below:

The error taxonomy is as follows:

1. Missing words (when one or more words are missing in the translation)
   - Missing word fillers
   - Missing content words
2. Word order (when reordering model is unable to produce a reordering of the sentence)
3. Incorrect words (when a translation engine is unable to produce a correct translation of a word or expression)
   - Lexical choice
   - Disambiguation
   - Incorrect form
   - Extra words
   - Idiomatic Expressions
4. Unknown words (when the translation engine could not find the translation in the target language and keeps the words or expressions in the source language).

Example extracted from the paper in Figure 8:

Original: *Who was the first American to walk in space?*
Translation: *Quem foi o primeiro a andar **americano** no espaço?*
Correct Translation: *Quem foi o primeiro **americano** a andar no espaço?*

Figure 8: Examples of word order error extracted from (Costa et al., 2012)

Macklovitch (1991) presents an error tabulation for its taxonomy in Figure 9.

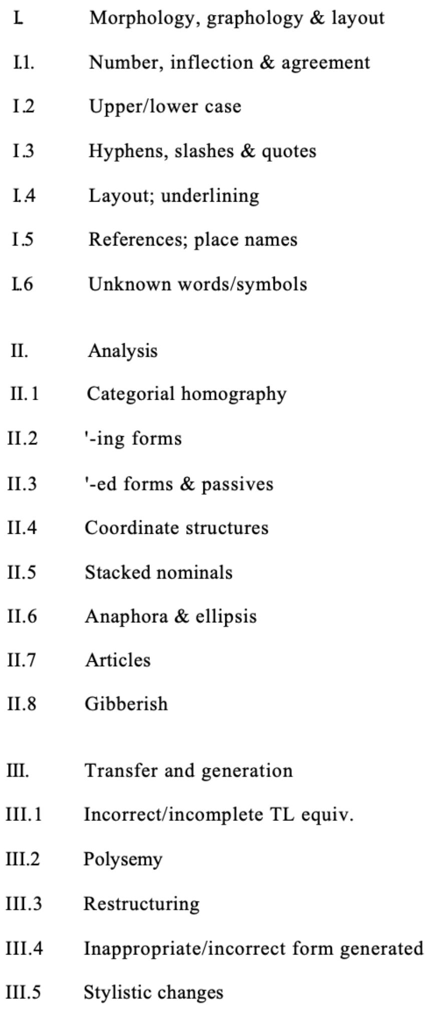
B Appendix: Error Annotation Schemes for Human Generated Text (HGT)

In this section, we present entire taxonomies and examples of individual errors for error annotation schemes developed for HGT and extracted from the 22 surveyed papers briefly summarised in Section 3.

Lin et al. (2022) present a taxonomy based on error correction and each edit belongs to one of the three types listed below. Figure 10 show error types and examples extracted from the paper.

The error taxonomy in Lin et al. (2022) is

1. Monolingual edits (identifiable from the target side of the text)
   - typos (spelling, punctuation, spacing and orthographic issues)
   - grammar



I. Morphology, graphology & layout
I.1. Number, inflection & agreement
I.2 Upper/lower case
I.3 Hyphens, slashes & quotes
I.4 Layout; underlining
I.5 References; place names
I.6 Unknown words/symbols

II. Analysis
II. 1 Categorial homography
II.2 '-ing forms
II.3 '-ed forms & passives
II.4 Coordinate structures
II.5 Stacked nominals
II.6 Anaphora & ellipsis
II.7 Articles
II.8 Gibberish

III. Transfer and generation
III.1 Incorrect/incomplete TL equiv.
III.2 Polysemy
III.3 Restructuring
III.4 Inappropriate/incorrect form generated
III.5 Stylistic changes

Figure 9: Figure of taxonomy extracted from (Macklovitch, 1991)

  - fluency

2. Bilingual edits (mismatch between source and target text) Eg, under-translation, mistranslation.
3. Preferential edits (based on the preference of the customer) Eg, terminology, style preference.

van der Goot et al. (2018) present a taxonomy; below are a few examples of errors.

Example 1: Typographical error

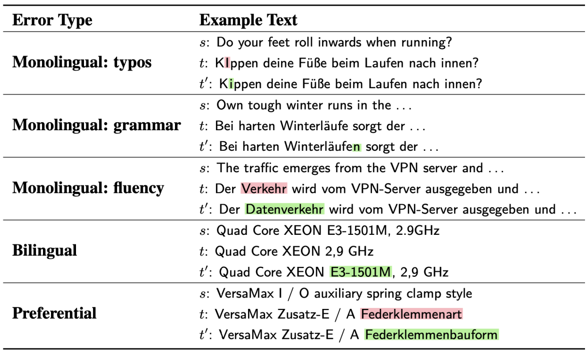
spirite|→spirit, complaing|→complaining, throwg|→throw

Example 2: Repetition

soooo|→so, weiiiiird|→weird

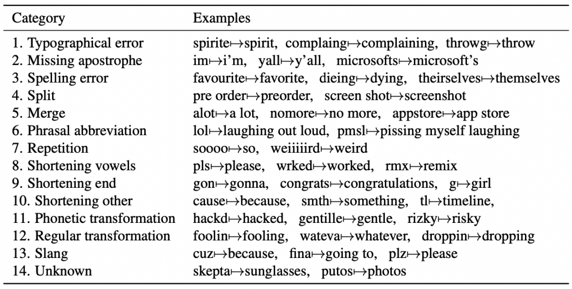
Example 3: Unknown

skepta|→sunglasses, putos|→photos



Error Type	Example Text
Monolingual: typos	s: Do your feet roll inwards when running? t: Klppen deine Füße beim Laufen nach innen? t': Kippen deine Füße beim Laufen nach innen?
Monolingual: grammar	s: Own tough winter runs in the ... t: Bei harten Winterläufe sorgt der ... t': Bei harten Winterläufen sorgt der ...
Monolingual: fluency	s: The traffic emerges from the VPN server and ... t: Der Verkehr wird vom VPN-Server ausgegeben und ... t': Der Datenverkehr wird vom VPN-Server ausgegeben und ...
Bilingual	s: Quad Core XEON E3-1501M, 2.9GHz t: Quad Core XEON 2,9 GHz t': Quad Core XEON E3-1501M, 2,9 GHz
Preferential	s: VersaMax I / O auxiliary spring clamp style t: VersaMax Zusatz-E / A Federklemmenart t': VersaMax Zusatz-E / A Federklemmenbauform

Figure 10: Figure of error taxonomy for ACED corpus with examples extracted from (Lin et al., 2022).



Category	Examples
1. Typographical error	spirite↦spirit, complaing↦complaining, throwg↦throw
2. Missing apostrophe	im↦i'm, yall↦y'all, microsofts↦microsoft's
3. Spelling error	favourite↦favorite, dieing↦dying, theirselves↦themselves
4. Split	pre order↦preorder, screen shot↦screenshot
5. Merge	alot↦a lot, nomore↦no more, appstore↦app store
6. Phrasal abbreviation	lol↦laughing out loud, pmsl↦pissing myself laughing
7. Repetition	soooo↦so, weiiiiird↦weird
8. Shortening vowels	pls↦please, wrked↦worked, rmx↦remix
9. Shortening end	gon↦gonna, congrats↦congratulations, g↦girl
10. Shortening other	cause↦because, smth↦something, tl↦timeline,
11. Phonetic transformation	hackd↦hacked, gentille↦gentle, rizky↦risky
12. Regular transformation	foolin↦fooling, wateva↦whatever, droppin↦dropping
13. Slang	cuz↦because, fina↦going to, plz↦please
14. Unknown	skepta↦sunglasses, putos↦photos

Figure 11: Figure of taxonomy extracted from (van der Goot et al., 2018).

Ng et al. (2014) present an error annotation scheme and a few examples below.

Figure 12 includes extracted error categories and its examples from the paper. Here are some examples extracted from the paper:

Example 1: Verb tense (Vt)

Medical technology during that time [is → was] not advanced enough to cure him.

Example 2: Word form (Wform)

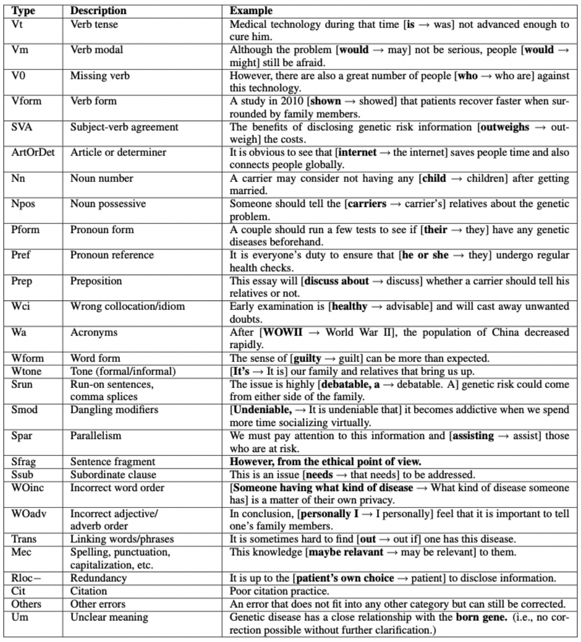
The sense of [guilty → guilt] can be more than expected.

Example 3: Unclear meaning (Um)

Genetic disease has a close relationship with the born gene. (i.e., no correction possible without further clarification.)

Dickinson and Herring (2008) presents a taxonomy for Russian verbal morphology:

1. Inappropriate verb stem
   - Always inappropriate



Type	Description	Example
Vt	Verb tense	Medical technology during that time [**is** → was] not advanced enough to cure him.
Vm	Verb modal	Although the problem [**would** → may] not be serious, people [**would** → might] still be afraid.
V0	Missing verb	However, there are also a great number of people [**who** → who are] against this technology.
Vform	Verb form	A study in 2010 [**shown** → showed] that patients recover faster when surrounded by family members.
SVA	Subject-verb agreement	The benefits of disclosing genetic risk information [**outweighs** → outweigh] the costs.
ArtOrDet	Article or determiner	It is obvious to see that [**internet** → the internet] saves people time and also connects people globally.
Nn	Noun number	A carrier may consider not having any [**child** → children] after getting married.
Npos	Noun possessive	Someone should tell the [**carriers** → carrier's] relatives about the genetic problem.
Pform	Pronoun form	A couple should run a few tests to see if [**their** → they] have any genetic diseases beforehand.
Pref	Pronoun reference	It is everyone's duty to ensure that [**he or she** → they] undergo regular health checks.
Prep	Preposition	This essay will [**discuss about** → discuss] whether a carrier should tell his relatives or not.
Wci	Wrong collocation/idiom	Early examination is [**healthy** → advisable] and will cast away unwanted doubts.
Wa	Acronyms	After [**WOWII** → World War II], the population of China decreased rapidly.
Wform	Word form	The sense of [**guilty** → guilt] can be more than expected.
Wtone	Tone (formal/informal)	[**It's** → It is] our family and relatives that bring us up.
Srun	Run-on sentences, comma splices	The issue is highly [**debatable, a** → debatable. A] genetic risk could come from either side of the family.
Smod	Dangling modifiers	[**Undeniable,** → It is undeniable that] it becomes addictive when we spend more time socializing virtually.
Spar	Parallelism	We must pay attention to this information and [**assisting** → assist] those who are at risk.
Sfrag	Sentence fragment	**However, from the ethical point of view.**
Ssub	Subordinate clause	This is an issue [**needs** → that needs] to be addressed.
WOinc	Incorrect word order	[**Someone having what kind of disease** → What kind of disease someone has] is a matter of their own privacy.
WOadv	Incorrect adjective/ adverb order	In conclusion, [**personally I** → I personally] feel that it is important to tell one's family members.
Trans	Linking words/phrases	It is sometimes hard to find [**out** → out if] one has this disease.
Mec	Spelling, punctuation, capitalization, etc.	This knowledge [**maybe relavant** → may be relevant] to them.
Rloc−	Redundancy	It is up to the [**patient's own choice** → patient] to disclose information.
Cit	Citation	Poor citation practice.
Others	Other errors	An error that does not fit into any other category but can still be corrected.
Um	Unclear meaning	Genetic disease has a close relationship with the **born gene.** (i.e., no correction possible without further clarification.)

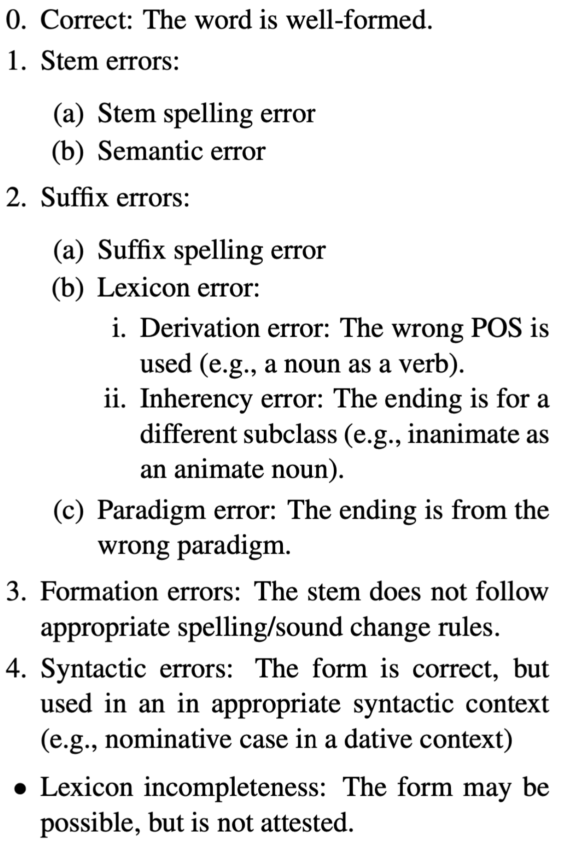
Figure 12: Figure of taxonomy extracted from (Ng et al., 2014).

- Inappropriate for this context

2. Inappropriate verb affix
   - Always inappropriate
   - Always inappropriate for verbs
   - Inappropriate for this verb
3. Inappropriate combination of stem and affix
4. Well-formed word in inappropriate context
   - Inappropriate agreement features
   - Inappropriate verb form (tense, perfective/imperfective, etc.)

Dickinson (2010) present an error taxonomy in Figure 13.
Nagata et al. (2018) present a spelling error annotation scheme in Figure 14.
Barbagli et al. (2016) present error annotations and examples in Figure 15.
Himoro and Pareja-Lora (2020) present an error taxonomy and examples in Figure 16.

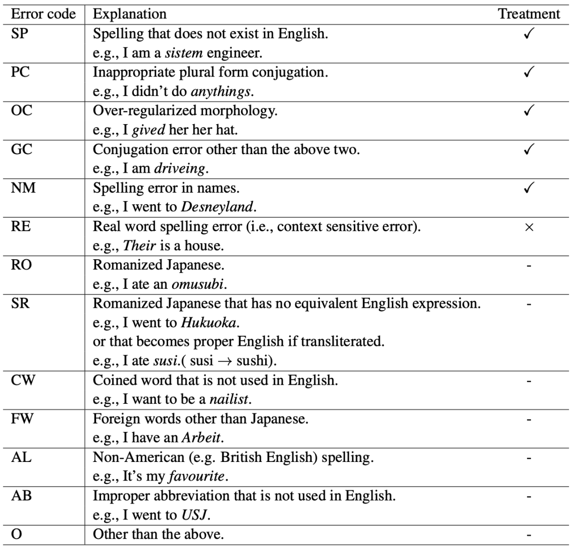
1. Abbreviation (ABR) in Intentional errors are due to omission of letters or use of homophones letters and/or numbers to replace syllables. Example, kme->kame.
2. Omission (OMS) in unintentional errors are due to deletion of letter from a word without an explanation. Example, Chaacano -> Chavacano.



0. Correct: The word is well-formed.
1. Stem errors:
   (a) Stem spelling error
   (b) Semantic error
2. Suffix errors:
   (a) Suffix spelling error
   (b) Lexicon error:
      i. Derivation error: The wrong POS is used (e.g., a noun as a verb).
      ii. Inherency error: The ending is for a different subclass (e.g., inanimate as an animate noun).
   (c) Paradigm error: The ending is from the wrong paradigm.
3. Formation errors: The stem does not follow appropriate spelling/sound change rules.
4. Syntactic errors: The form is correct, but used in an in appropriate syntactic context (e.g., nominative case in a dative context)

- Lexicon incompleteness: The form may be possible, but is not attested.

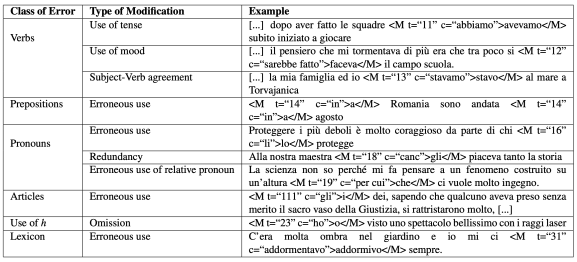
Figure 13: Figure of taxonomy extracted from (Dickinson, 2010)



Error code	Explanation	Treatment
SP	Spelling that does not exist in English. e.g., I am a *sistem* engineer.	✓
PC	Inappropriate plural form conjugation. e.g., I didn't do *anythings*.	✓
OC	Over-regularized morphology. e.g., I *gived* her her hat.	✓
GC	Conjugation error other than the above two. e.g., I am *driveing*.	✓
NM	Spelling error in names. e.g., I went to *Desneyland*.	✓
RE	Real word spelling error (i.e., context sensitive error). e.g., *Their* is a house.	×
RO	Romanized Japanese. e.g., I ate an *omusubi*.	-
SR	Romanized Japanese that has no equivalent English expression. e.g., I went to *Hukuoka*. or that becomes proper English if transliterated. e.g., I ate *susi*.(susi → sushi).	-
CW	Coined word that is not used in English. e.g., I want to be a *nailist*.	-
FW	Foreign words other than Japanese. e.g., I have an *Arbeit*.	-
AL	Non-American (e.g. British English) spelling. e.g., It's my *favourite*.	-
AB	Improper abbreviation that is not used in English. e.g., I went to *USJ*.	-
O	Other than the above.	-

Figure 14: Figure of spelling error and corresponding treatment extracted from (Nagata et al., 2018).

Caines et al. (2020) present the error types determined by grammatical error correction of texts in TSCC in Figure 17.
Korre et al. (2021) present an error annotation scheme in Figure 18.



Class of Error	Type of Modification	Example
Verbs	Use of tense	[...] dopo aver fatto le squadre <M t="11" c="abbiamo">avevamo</M> subito iniziato a giocare
	Use of mood	[...] il pensiero che mi tormentava di più era che tra poco si <M t="12" c="sarebbe fatto">faceva</M> il campo scuola.
	Subject-Verb agreement	[...] la mia famiglia ed io <M t="13" c="stavamo">stavo</M> al mare a Torvajanica
Prepositions	Erroneous use	<M t="14" c="in">a</M> Romania sono andata <M t="14" c="in">a</M> agosto
Pronouns	Erroneous use	Proteggere i più deboli è molto coraggioso da parte di chi <M t="16" c="li">lo</M> protegge
	Redundancy	Alla nostra maestra <M t="18" c="canc">gli</M> piaceva tanto la storia
	Erroneous use of relative pronoun	La scienza non so perché mi fa pensare a un fenomeno costruito su un'altura <M t="19" c="per cui">che</M> ci vuole molto ingegno.
Articles	Erroneous use	<M t="111" c="gli">i</M> dei, sapendo che qualcuno aveva preso senza merito il sacro vaso della Giustizia, si rattristarono molto, [...]
Use of *h*	Omission	<M t="23" c="ho">o</M> visto uno spettacolo bellissimo con i raggi laser
Lexicon	Erroneous use	C'era molta ombra nel giardino e io mi ci <M t="31" c="addormentavo">addormivo</M> sempre.

Figure 15: Figure of error annotations and examples extracted from (Barbagli et al., 2016)

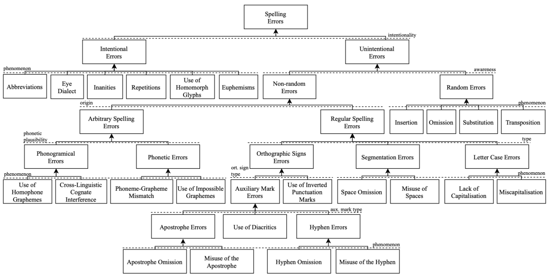
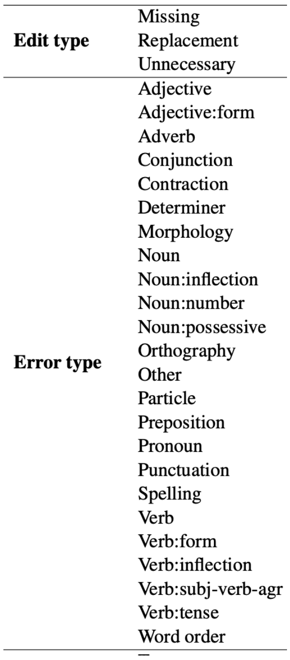
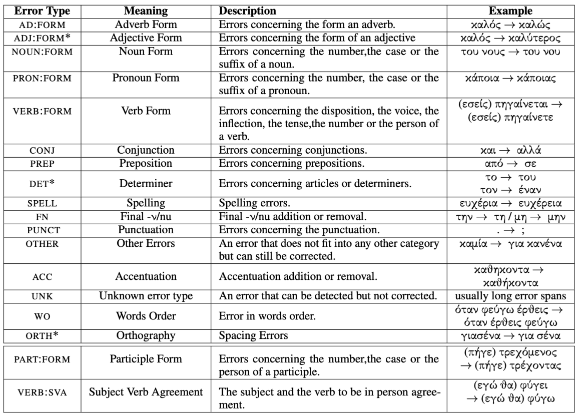


Figure 16: Figure of spelling error taxonomy for ZC extracted from (Himoro and Pareja-Lora, 2020)



Edit type	Missing Replacement Unnecessary
Error type	Adjective Adjective:form Adverb Conjunction Contraction Determiner Morphology Noun Noun:inflection Noun:number Noun:possessive Orthography Other Particle Preposition Pronoun Punctuation Spelling Verb Verb:form Verb:inflection Verb:subj-verb-agr Verb:tense Word order

Figure 17: Figure of the error types determined by grammatical error correction of texts in TSCC extracted from (Caines et al., 2020)



Error Type	Meaning	Description	Example
AD:FORM	Adverb Form	Errors concerning the form an adverb.	καλός → καλώς
ADJ:FORM*	Adjective Form	Errors concerning the form of an adjective	καλός → καλύτερος
NOUN:FORM	Noun Form	Errors concerning the number,the case or the suffix of a noun.	του νους → του νου
PRON:FORM	Pronoun Form	Errors concerning the number, the case or the suffix of a pronoun.	κάποια → κάποιας
VERB:FORM	Verb Form	Errors concerning the disposition, the voice, the inflection, the tense,the number or the person of a verb.	(εσείς) πηγαίνεται → (εσείς) πηγαίνετε
CONJ	Conjunction	Errors concerning conjunctions.	και → αλλά
PREP	Preposition	Errors concerning prepositions.	από → σε
DET*	Determiner	Errors concerning articles or determiners.	το → του τον → έναν
SPELL	Spelling	Spelling errors.	ευχέρια → ευχέρεια
FN	Final -ν/nu	Final -ν/nu addition or removal.	την → τη / μη → μην
PUNCT	Punctuation	Errors concerning the punctuation.	. → ;
OTHER	Other Errors	An error that does not fit into any other category but can still be corrected.	καμία → για κανένα
ACC	Accentuation	Accentuation addition or removal.	καθηκοντα → καθήκοντα
UNK	Unknown error type	An error that can be detected but not corrected.	usually long error spans
WO	Words Order	Error in words order.	όταν φεύγω έρθεις → όταν έρθεις φεύγω
ORTH*	Orthography	Spacing Errors	γιασένα → για σένα
PART:FORM	Participle Form	Errors concerning the number,the case or the person of a participle.	(πήγε) τρεχόμενος → (πήγε) τρέχοντας
VERB:SVA	Subject Verb Agreement	The subject and the verb to be in person agreement.	(εγώ θα) φύγει → (εγώ θα) φύγω

Figure 18: Figure of ELERRANT and human error type annotation guide extracted from (Korre et al., 2021). The error types with (*) do not exist for human annotation scheme and the last two error types do not exist in the ELERRANT annotation scheme.