SMT error analysis and mapping to syntactic, semantic and structural fixes

Nora Aranberri
IXA Group
University of the Basque Country
Manuel Lardizabal 1, 20018 Donostia, Spain
nora.aranberri@ehu.eus

Abstract

This paper argues in favor of a linguistically-informed error classification for SMT to identify system weaknesses and map them to possible syntactic, semantic and structural fixes. We propose a scheme which includes both linguistic-oriented error categories as well as SMT-oriented edit errors, and evaluate an English-Spanish system and an English Basque system developed for a Q&A scenario in the IT domain. The classification, in our use-scenario, reveals great potential for fixes from lexical semantics techniques involving entity handling for IT-related names and user interface strings, word sense disambiguation for terminology, as well as argument structure for prepositions and syntactic parsing for various levels of reordering.

1 Introduction

Once we build a baseline SMT system, we run an evaluation to check its performance and guide improvement. Given the nature of statistical systems and their learning process, linguistic-oriented error analysis has been considered unfit for their evaluation. Even when it is identified that a particular linguistic feature is incorrectly handled, it is not clear how to specifically address it during training if we resort to common generic, non-deterministic techniques. However, when syntax, semantics and structure (SSS) come into play, error analysis regains relevance, as it can pinpoint specific aspects that can be addressed through the more targeted techniques they have brought to MT development.

Based on two baseline SMT systems, one for the English-Spanish pair and one for English-Basque, we present a methodology and classification for error analysis, a description of the results and a mapping to possible fixes using SSS techniques.

2 Error classification schemes

Different classification schemes have been proposed in the last years to categorize machine translation errors. Starting in the 90s, the LISA QA model was adopted by good part of the industry.[1] This model included a list of "objective" error types, graded by their severity and pre-assigned penalty points. The SAE J2450 standard, from the automotive service, also became popular.[2] What became clear from these first efforts was that no one-fits-all evaluation scheme is possible for MT. Each player within the translation workflow, from developers to vendors and clients, has its own needs and the information they expect from the evaluations is different.

After LISA ceased operations, two major efforts emerged: TAUS presented its Dynamic Quality Framework (DQF) [3] and the QTLaunchPad project developed the Multidimensional Quality Metrics (MQM).[4] The DQF tackles quality evaluation by identifying the objective of each evaluation and by offering a bundle of tools to satisfy each need. Specifically, they offer productivity testing based on post-editing effort, adequacy and fluency tests,

[1] The Localization Industry Standards Association terminated activities in 2011. No official reference is now available.

[2] SAE J2450: http://www.sae.org/standardsdev/j2450p1.htm

[3] DQF: https://evaluate.taus.net

[4] MQM: http://www.qt21.eu/mqm-definition

Proceedings of SSST-9, Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 30–38,
Denver, Colorado, June 4, 2015.

translation comparisons and error classification. The error scheme, proposed after a thorough examination of industry practices, covers four main areas, namely, *Accuracy*, *Language*, *Terminology* and *Style*, with limited subcategories. With a strong industrial view, it focuses on establishing return-on-investment and on benchmarking performance to allow for informed decisions, rather than providing a detailed development-oriented error analysis.

The MQM is a framework that can be used to define metrics in order to assign a level of quality to a text. Each evaluation must identify the relevant categories for its goals and *customize* the metric. MQM Core is a hierarchy of 22 issues, at different levels of granularity. If we consider *Accuracy* and *Fluency*, the two top-level categories that best focus on intratextual diagnosis, subcategories branch out and get more detailed, although they remain at a relatively general level. Authors claim that considerably more detailed subclasses might be necessary to diagnose MT problems and the framework allows for user-defined extensions, even if this is not encouraged.

The MQM puts together three different dimensions of error classification. The two top-level categories, *Accuracy* and *Fluency*, can be seen as the effect the errors have on a translated text. The concepts in the lower-levels include concepts of yet another two dimensions. Some of the subcategories refer to actual errors systems make, such as mistranslation or grammar, whereas others refer to the way in which these errors are rendered, namely, omission, addition and incorrect. When trying out the scheme to perform our evaluation, we saw that the distinction between fluency and accuracy might, to some extent, be useful when prioritizing fixes. However, we found difficulty in assigning an error to a specific subclass, as overlaps between dimensions occurred constantly. For example, grammar is placed under fluency but we could argue that an incorrect tense might lead to a significant change in meaning, and therefore, result in an accuracy issue. Similarly, one could claim that the rendering possibilities are true for almost, if not all, types of errors, rather than a category of their own. For example, Addition is a direct subclass of Accuracy, even if it is possible to find extra function words in a translation. Also, we strongly felt that some subclasses were too broad to be meaningful to decide on a targeted SSS solution.

Among schemes that have emerged from research groups, Vilar et al. (2006) presented one of the first to focus on identifying errors made by statistical systems. Probably motivated by the fact that these systems are not controlled by linguistic rules and are not deterministic in this respect, the top-level categories proposed were *Missing words*, *Word order*, *Incorrect words*, *Unknown words* and *Punctuation*, that is, types of edits unrelated to linguistic reasoning. The lower categories are slightly more linguistic but they remain on SMT parameters such as *local/long range*, *stems* and *forms*. While *Word order* and *Unknown words* point to specific efforts for improvement, the *Incorrect words* category is broad and requires, as the authors suggest, further customization depending on the language pair at hand. Again, this classification lacks the linguistic detail we aimed to collect for linguistically-oriented fixes.

2.1 Classification schemes: our approach

Given our goal and the nature of our systems, we opted for a general linguistic classification with an additional dimension to cover the edit type of each error: missing, additional or incorrect (Figure 1). Once a linguistic error is identified, it is classified based on the edit-type dimension. We established six top-level linguistic categories, which are further detailed in subclasses. These subclasses are not static but rather they can be omitted or extended during evaluation to suit errors found in texts. The linguistic depth and the clear division between dimensions overcomes the lack of detail of the DQF model and the overlaps that emerged in the MQM model, while incorporating the SMT-oriented edits proposed by Vilar et al. (2006).

We worked with a two-to-four-level scheme to gather as much detail as possible about the errors found. We describe the six main categories below.

Top-level category	Subclasses	Incorrect	Missing	Additional
Lexis				
Morphosyntax				
Verbs				
Order				
Punctuation				
Untranslated				

Figure 1: Proposed bidimensional error scheme.

1. Lexis

   This category includes incorrect choices for general vocabulary and terminology, as well as longer set phrases, idioms or expressions.

2. Morphosyntax

   This category includes morphological and syntactic errors. We fused both categories as these types of errors are often so intertwined that it is difficult to opt for one category over the other. Moreover, the classification is proposed as a tool to easily summarize and assimilate system error information and the exact top-level classification of the items should not have an impact on research decisions. This should be guided by their fixing requirements and possibilities.

3. Verbs

   A separate category was defined for verb phrases because of their complexity. Whereas English verb phrases carry lexical, aspectual, tense, modality and voice information, Spanish verb phrases also have subject information, and in the case of Basque, information about objects is also included. The high variability of conjugated verbs and auxiliaries poses great difficulty for statistical systems. We divided this category into subgroups based on the information mentioned above.

4. Order

   Again, this is a dedicated category due to the impact order has on the overall comprehensibility of the translations and because it is a property that can be addressed specifically in statistical systems. We distinguished several levels: sentence, clause and phrase. Also, we identify whether the issues involve orderings of units of the same level or, unit-specific issues, which can be internal orderings or splits.

5. Punctuation

   This category includes punctuation and orthographic issues such as punctuation marks, capitalization and orthotactic constrains (orthographic rules governing lemma-affix gluing).

6. Untranslated

   We added a category for source words that are left in the original language.

3 The systems

3.1 English-Spanish

The English-Spanish system is a standard phrase-based system built on Moses (Koehn, et al. 2007). It uses basic tokenization and a pattern excluding URLs, truecasing and language model interpolation. It has been trained on bilingual corpora including Europarl, United Nations, News Commentary and Common Crawl (∼355 million words). The monolingual corpora used to learn the language model include the Spanish texts of Europarl, News Commentary and News Crawl (∼60 million words). For tuning, a set of 1,000 in-domain interactions (question-answer pairs) were made available. The original interactions are in English and they were translated into Spanish by human translators.

The system was evaluated on a test-set similar to that used for tuning: a second batch of 1,000 in-domain interactions. The English-Spanish system obtains a BLEU score of 45.86.

3.2 English-Basque

The English-Basque system is also a standard phrase-based system built on Moses. It uses basic tokenization, lemmatization and lowercasing. Stanford CoreNLP (Manning et al., 2014) is used for English analysis and Eustagger (Alegría et al., 2002) for Basque. It uses a 5-gram language model. To better address the agglutinative nature of Basque, the word alignments were obtained over the lemmas, and were then projected to the original word forms to complete the training process.

The system was trained on translation memory (TM) data containing academic books, software manuals and user interface strings (∼12 million words), and web-crawled data (∼1.5 million words) made available by Elhuyar.[5] For the language model, the Basque text of the parallel data and the Basque text of Spanish-Basque TMs of administrative text made available by Elhuyar (∼7.4 million sentences) was used. Again, a set of 1,000 in-domain interactions were used for tuning after manually translating the original text into Basque.

The system was evaluated on a second test-set of 1,000 in-domain interactions, obtaining a BLEU score of 20.24.

[5] Elhuyar: https://www.elhuyar.eus/en

Error category	Examples
lexis	Click *run* where it says vulnerabilities. Pulse *correr* donde dice vulnerabilidades. (run=sport)
morphosyntax	Yes, you can share files and folders with one or more users *on MEO* Cloud. Sí, puede compartir archivos y carpetas con uno o más usuarios *sobre* MEO Cloud. (on=about)
verb	*Connect* your computer to the ZON HUB via Ethernet cable. *Conectar* su ordenador a la HUB af a travs de cable Ethernet. (to connect)
ordering	Tap "Import" to copy your *Android browser favorites*. Toca "Importar" para copiar su *navegador de Android favoritos*. (∼your favorites Android browser)
punctuation	If I buy a computer abroad, will it work in Portugal Si compro un ordenador en el extranjero, funcionará en Portugal*?* (missing ¿)
untranslated	*Then* click on the yellow disc with a green tick. *Then* haga clic en el disco de color amarillo con una marca verde.

Table 1: Examples of errors per top-level category for the English-Spanish pair.

4 Error analysis results [6]

4.1 Error analysis for the English-Spanish pair

We randomly selected 100 interventions (questions or answers) included in the use-scenario test set. Overall, out of 137 sentences (each intervention might consist of several sentences) 30 sentences were found to be correct, and the remaining 107 include 169 errors, at least 3 errors per intervention.

Lexical errors account for 31% of the total mistakes (see examples for top-level categories in Table 1). Around half emerge from the translation of user interface (UI) strings. Although it was not possible to identify whether the translations matched the final software version text exactly, in some cases the translations are clearly awkward. Problems are most relevant in multi-word strings, which are not translated as a unit, resulting in partial translations and inadequate capitalization. The translations of software and brand names display a similar behavior. These proper names tend to stay the same across languages, but the system does not always treat them this way. Adding to this, multiword names often get part of the name translated.

Issues with general vocabulary and terminology (we will consider terminology words that acquire a specialized meaning in our domain or words that are specific to our domain) are also present. Whereas some inadequate translations do not have a clear origin, a good number of them clearly emerge from incorrect word sense disambiguation.

Morphosyntactic errors account for about 29% of the total errors. Although they are very widespread across the different subcategories, we find that prepositions, subordinate markers and POS errors are the most recurrent cases.

The Verbs category accounts for 18% of the errors. Although a number of verbs lack the correct agreement or use an inadequate tense or voice, the most recurrent error seems to come from the mode. This is typical of instructional texts, where orders, given with the infinitive form in English can be translated as imperatives or infinitives. This is usually a stylistic decision but one that needs to be consistent across the documentation and, in particular, within the sentence or paragraph.

A number of order issues have been identified (11%), which mainly involve the composition of multiword noun phrases. We found 7 cases where a noun phrase was split and 7 cases where the elements were incorrectly ordered despite staying in close proximity.

Punctuation errors (6%) and untranslated words (5%) are low. The former include cases of incorrect capitalization and use of question-initial marks. The latter involve function and content words.

4.2 Error analysis for the English-Basque pair

We again performed a random selection of 100 interventions. Based on overall counts, 6 out of 140 sentences were correct and the remaining 134 included 393 errors, at least 7 errors per intervention.

Lexical errors account for around 23% of the total (Table 2). Despite a number of errors due to incorrect word sense disambiguation, most errors emerge from UI strings and software/brand name translations. Capitalization errors in these units were included in this subcategory (36 cases).

Morphosyntactical errors account for over 39%

[6] For a complete classification see appendices A and B.

Error category	Examples
lexis	Go to WhatsApp > *"Menu Button"* > "Status". Joan *menu botoia* WhatsApp > " " > " egoera ". (unrecognized user interface path)
morphosyntax	Yes it is possible, simply by dragging the profile of the person concerned *to* the various circles. Bai posible da, besterik gabe, arrastatu pertsonaren profila hainbat nahia zirkulu. (missing postposition for circles-zirkulu)
verb	Choose a standard status or personalize one. Egoera estandar bat edo pertsonalizatu bat. (missing verb choose)
ordering	*You can use the app iPP Podcast Player* you find on Google Play. *Aplikazioa erabil dezakezu IPP podcast Player* aurkitu duzu Google erreproduzitu. (The app you can use IPP podcast Player...)
punctuation	How can I change the language to of Mega to Portuguese? Nola aldatu hizkuntza of Mega, portugesa? (additional comma)
untranslated	How much space do I have for free *on* Mega? Zenbat leku ditut doan *on* Mega?

Table 2: Examples of errors per top-level category for the English-Basque pair.

of the total errors. Most, around 64%, concern the translation of prepositions and subordinate conjunctions. In Basque, prepositions are translated into postpositions that are attached to the last word of the phrase (the nucleus) and the same happens with subordinate markers, attached to the last word of the subordinate clause. It is worth noting the high number of missing elements in this subcategory, 90 cases recorded out of 149 (10 cases out of 49 for Spanish).

Verbs show a considerable number of errors (18%), specially if we take into account that 21 main verbs, which display the lexical meaning and the aspect, and 23 auxiliaries, which display tense, mode and paradigm, are missing. Out of the verb phrases that are constructed, the aspect, the paradigm and agreements generate errors.

Order errors account for 14% of the total errors. The sequencing of noun phrase elements stands out as the main source of errors, whether within the phrase or because splits occurred. The positioning of relative clauses with respect to their heads also emerged as a problematic area with 11 occurrences.

Punctuation (4%) and untranslated words (1%) are low, the most salient being missing commas.

4.3 Fixing possibilities with syntax, semantics and structure

From the error analysis of the English-Spanish and English-Basque systems we see that errors emerge from two main sources, use-scenario-specific features and language pair-specific features.

The text-type and domain of the translations has an impact on the difficulties the system encounters. In the case we present, we work on a question-and-answer (Q&A) scenario in the information technology (IT) domain. The texts, therefore, mainly consist of instructions and descriptions, and include a high degree of terminology, brand and software names, as well as UI strings. And our systems have difficulty in dealing with them.

Lexical semantics, and in particular, (cross-lingual) named-entity recognition (NER) and translation techniques could greatly benefit our application scenario. Following the implementation of NER in MT by Li et al. (2013), Li et al. (2012) and similar, it would be possible to train a NER system to identify IT names. We could possibly create a separate category for the disambiguation process (NED) if we envisage to treat them in a specific way. For example, we may decide that NEs classified as *IT-name* should be left in English, or that they should be looked up in Wikipedia following techniques such as Mihalcea and Csomai's (2007) and Agirre et al.'s (2015) to find an equivalent entry in the target language, and as a result, its translation. Maybe we could opt for dynamic searches in multilingual websites of specific brands or the use of pre-compiled dictionaries from these resources.

The NER system could be expanded to include UIs. Cues to identify them could be anchors like *icon, tab* and *dialog box*, and phrases such as *where it says*, and > sequences. The systems had difficulty in identifying UIs and often provided translations that differ significantly from the strings we are used to seeing in software graphics. UIs usually have a fixed translation - often given by the product-maker - and they must be treated as proper nouns in the sense that they are usually capitalized (first word only if multiword) and do not accept articles. We could chose to identify them and translate them us-

ing a specialized dictionary or even let the MT system output a candidate which considers the restrictions just mentioned.

Sense disambiguation, whether for general words or terms, has also been identified as a category worth addressing. Word sense disambiguation techniques along the line of Carpuat et al. (2013), for example, could help. They propose a technique to identify unknown senses to the system, most probably because they are domain-specific senses not covered by the training corpus. Once marked, we could divert them and translate them using a specialised resource.

Out of the language pair-specific errors, the most glaring are Basque postpositional renderings of English prepositions. Predicate-argument structures and semantic roles, as suggested by the work of Liu and Gildea (2010) and Kawahara and Kurohashi (2010), are a way to improve the incorrect renderings and to force missing postpositions. Resources such as the Basque Verb Index (BVI) (Estarrona et al., forthcoming), which includes Basque verb subcategorization based on PropBank and VerbNet, with syntactic renderings assigned to each argument and mappings to WordNet for crosslingual information, can be a starting point in this task.

Order errors have shown three types of issues: (i) phrases or chunks ordered incorrectly; (ii) phrases split along the sentence; and (iii) phrasal elements kept local but with incorrect phrase-internal order. For the first case, semantics has proposed the use of argument structure to learn reordering patterns (Wu et al., 2011). For cases ii and iii, syntax would have to come into play. Firstly, we need to provide the MT with phrase boundary information so that contiguous phrases are not mixed. Secondly, phrase-internal reordering patterns or restrictions need to apply. Yeniterzi and Oflazer (2010), for example, encode a variety of local and non-local syntactic structures of the source side as complex structural tags and include this information as additional factors during training. Also, working on POS, Popović and Ney (2006) propose source-side local reordering patterns for Spanish-English and, working on syntactic parse-level, Wang et al. (2007) propose reordering patterns to address systematic differences (Chinese-English). Xiong et al. (2010) go beyond syntax and propose translation zones as unit boundaries, improving constituent-based approaches.

We finally focus on the generation of verb phrases, particularly relevant for the English-Basque pair, where verbs tend to go missing, but also to remedy incorrect verbal features in both pairs. The sparsity due to the complexity and morphological variety of Spanish and, even more so, Basque verb phrases is most probably the main reason for their incorrect handling. This leads us to proposing the generalization of features, such as lemmatization of verbs, while suggesting a parallel transfer of source verb features to final postprocessing, for instance. Work on verbal transfer has not received attention so far, unless integrated within argument structure techniques, such as the work of Xiong et al. (2012).

5 Conclusions

We proposed a dynamic, extensible linguistically-informed error classification for SMT which includes six top-level linguistic error categories with further subclasses, and a second dimension for SMT-oriented edits covering additions, omissions and incorrect words. This addresses the lack of linguistic detail and flexibility of metrics such as the DQF, and integrates the SMT-oriented errors proposed by Vilar et al. (2006) avoiding overlaps found in MQM.

We evaluated an English-Spanish and an English-Basque system developed for a Q&A scenario in the IT domain. The classification revealed issues strongly related to the domain and more general language pair-specific errors. We identified terminology and UI strings as the main issue for the lexical category. The morphosyntactic category showed more diverging issues. The most striking was the weak handling of English prepositions, and in particular, the poor generation of Basque postpositions, governing English prepositions and subordinate markers. The complexity of target-side verbs also took its toll on system performance with incorrect features for Spanish and an alarming number of missing main verbs and auxiliaries for Basque. As expected, ordering errors occurred at all levels, internal and external. Punctuation and Untranslated showed a low number of errors.

The exercise served to link the potential relevance of syntax, semantics and structure to fix language-specific SMT errors and the suitability of lexical semantics for IT-domain terminology and UI strings.

Acknowledgments

The research leading to these results has received funding from FP7-ICT-2013-10-610516 (QTLeap).

References

Eneko Agirre, Ander Barrena and Aitor Soroa. 2015. *Studying the Wikipedia Hyperlink Graph for Relatedness and Disambiguation.* arXiv:1503.01655.

Iñaki Alegría, María Jesús Aranzabe, Anton Ezeiza, Nerea Ezeiza and Ruben Urizar. 2002. *Robustness and customisation in an analyser/lemmatiser for Basque.* Proceedings of the LREC-2002 Workshop on Customizing knowledge in NLP applications.

Marine Carpuat, Hal Daumé III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi and Rachel Rudinger. 2013. *Studying the Wikipedia Hyperlink Graph for Relatedness and Disambiguation.* Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, pages 1435–1445.

Ainara Estarrona, Izaskun Aldezabal, Aranza Díaz de Ilarraza and María Jesús Aranzabe. Forthcoming. *Methodology of construction of the corpus-based Basque Verb Index (BVI) Lexicon* Language Resources and Evaluation.

Daisuke Kawahara and Sadao Kurohashi. 2010. *Acquiring Reliable Predicate-argument Structures from Raw Corpora for Case Frame Compilation* Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valleta, Malta, pages 1389–1393.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin, Evan Herbst. 2007. *AMoses: open source toolkit for statistical machine translation.* Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Prague, Czech Republic.

Qi Li, Haibo Li, Heng Ji, Wen Wang, Jing Zheng and Fei Huang. 2012. *Joint Bilingual Name Tagging for Parallel Corpora.* In Proceeding of CIKM12, pages 1727–1731.

Haibo Li, Jing Zheng, Heng Ji, Qi Li and Wen Wang. 2013. *Name-aware Machine Translation* Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, pages 604–614.

Ding Liu and Daniel Gildea. 2010. *Semantic role features for machine translation* Proceedings of the 23rd International Conference on Computational Linguistics, pages 716–724.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard and David McClosky. 2014. *The Stanford CoreNLP Natural Language Processing Toolkit.* In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60.

Rada Mihalcea and Andras Csomai. 2007. *SenseSpotting: Never let your parallel data tie you to an old domain.* In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pages 233–242. ACM.

Maja Popović and Hermann Ney. 2006. *POS-based Word Reorderings for Statistical Machine Translation.* Proceedings on the fifth international conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, pages 1278–1283.

David Vilar, Jia Xu, Luis Fernando DHaro and Hermann Ney. 2006. *Error Analysis of Statistical Machine Translation Output.* Proceedings on the fifth international conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, pages 697–702.

Chao Wang, Michael Collins and Philipp Koehn. 2007. *Chinese Syntactic Reordering for Statistical Machine Translation.* Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, pages 737745.

Xianchao Wu, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Masaaki Nagata. 2011. *Extracting Preordering Rules from Predicate-Argument Structures.* Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, pages 29–37.

Deyi Xiong, Min Zhang and Haizhou Li. 2012. *Verb Translation and Argument Reordering.* Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, pages 902–911.

Deyi Xiong, Min Zhang and Haizhou Li. 2010. *Learning Translation Boundaries for Phrase-based Decoding.* Proceedings of the 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, California, pages 136–144.

Reyyan Yeniterzi and Kemal Oflazer. 2010. *Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish* Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, pages 454–464.

A Error classification scheme and results for the English-Spanish pair

Main category	Subcategory 1	Subcategory 2	Incorrect		Missing	Additional
Lexis (53)	Vocabulary	lexical choice	2	(53)		
		sense	6			
	Terminology	lexical choice	4			
		sense	6			
	software brand names		8			
	UI issues		27			
Morphosyntax (49)	POS		9	(31)	(10)	(8)
	preposition		7		1	3
	noun				1	2
	adjective	agreement	2			
	determiner	other			3	
		agreement	1			
	article					2
	pronoun	other	1			
		gender	1			
		formal vs informal	3			
	interrogative pronoun		3			
	attribute	other	1			
		agreement			1	
	coordinator					1
	subordinate markers	relative marker	1		2	
		completive marker			2	
		purpose marker	1			
	coreference	agreement	1			
Verbs (30)	verb phrase		2	(26)	(4)	
	subject agreement		1		2	
	tense		2		1	
	mode	other			1	
		disagreement	7			
		infinite vs imperative	9			
	voice	passive	3			
	auxiliary		2			
Order (19)	noun phrase - internal		9	(19)		
	split noun phrase		7			
	split prepositional phrase		1			
	verb-adverb		2			
Punctuation (10)	capitalization	uppercase	4	(4)	(5)	(1)
	accent				1	
	question mark				4	1
Untranslated (8)			8	(8)		
Total (169)				141	19	9

B Error classification scheme and results for the English-Basque pair

Main category	Subcategory 1	Subcategory 2	Incorrect		Missing		Additional	
Lexis (93)		lexical choice	3	(93)				
	Vocabulary	sense	3					
	Terminology	sense	4					
	software brand names		28					
	UI issues		55					
Morphosyntax (154)	POS		11	(51)		(90)		(11)
	preposition		23		43		2	
	noun	other			1			
		agreement	3					
	adjective			2	2		1	
	determiner				1		1	
	article		5		1			
	adverb				5			
	pronoun				1			
	interrogative pronoun		1		7		1	
	negation (verbs)				1		2	
	coordinator						1	
	coordinated subclause		5		4			
	superlative structure		2					
	subordinate markers	relative marker			11			
		relative marker			3		1	
		completive marker			3			
		purpose marker	1		4		2	
		reason marker			1			
		temporal marker			1			
		conditional marker			1			
Verbs (70)	verb phrase		4	(19)		(45)	2	(2)
	main verb					21		
	auxiliary verb					23		
	subject agreement			3				
	direct object agreement			2				
	tense			1				
	aspect		5					
	auxiliary			4		1		
	paradigm			4				
Order (55)	constituent-level		2	(55)				
	noun phrase - internal		19					
	split noun phrase		7					
	split prepositional phrase		3					
	clause-level		1					
	clause internal		2					
	clause split		1					
	head-relative clause		11					
	contiguous sentences merged		9					
Punctuation (16)	capitalization		2	(4)		(11)		(1)
	comma		2		11			
	EOS						1	
Untranslated (5)			5	(5)				
Total (393)				233		135		25