Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2685–2697
Marseille, 11–16 May 2020

Towards a Spell Checker for Zamboanga Chavacano Orthography

Marcelo Yuji Himoro*, **Antonio Pareja-Lora***,**
* ATLAS (UNED – Universidad Nacional de Educación a Distancia),
** Universidad Complutense de Madrid (UCM)
São Paulo, Brazil / Madrid, Spain
mhimoro1@alumno.uned.es, aplora@ucm.es

Abstract

Zamboanga Chabacano (ZC) is the most vibrant variety of Philippine Creole Spanish, with over 400,000 native speakers in the Philippines (as of 2010). Following its introduction as a subject and a medium of instruction in the public schools of Zamboanga City from Grade 1 to 3 in 2012, an official orthography for this variety - the so-called "Zamboanga Chavacano Orthography" - has been approved in 2014. Its complexity, however, is a barrier to most speakers, since it does not necessarily reflect the particular phonetic evolution in ZC, but favours etymology instead. The distance between the correct spelling and the different spelling variations is often so great that delivering acceptable performance with the current *de facto* spell checking technologies may be challenging. The goals of this research is to propose i) a spelling error taxonomy for ZC, formalised as an ontology and ii) an adaptive spell checking approach using Character-Based Statistical Machine Translation to correct spelling errors in ZC. Our results show that this approach is suitable for the goals mentioned and that it could be combined with other current spell checking technologies to achieve even higher performance.

Keywords: Chavacano, Chabacano, Zamboanga, Natural Language Processing, Under-resourced Languages, spell checker.

1. Introduction

According to a recent report from Komisyon sa Wikang Filipino (KWF, 2015), the Philippines is home to 135 living languages, among which Chabacano is the only creole language. In the 2010 Census of Population and Housing (CPH), 405,798 people claimed Zamboanga Chabacano (ZC, locally known simply as "Chavacano" or "Zamboangueño") as their mother tongue, thus making it the most widely spoken variety of Philippine Creole Spanish in the country, and the only one still experiencing natural growth (National Statistics Office (NSO), 2003a, 2014a; Philippine Statistics Authority (PSA), 2014).

With the implementation of the K-12 curriculum and the Mother-Tongue Based Multilingual Education (MTB-MLE) program nationwide in 2012, ZC is nowadays taught as a subject and is used as a medium of instruction in the public schools of Zamboanga City from Grade 1 to 3 (Government of the Philippines, 2011). This has motivated the local government to invest in the standardisation of ZC, resulting in the approval of an official orthography (Zamboanga Chavacano Orthography) in 2014, followed by the publication of a basic grammar, as well as some collections of texts and children's books; even a normative dictionary was published in late 2018.

Despite all the official initiatives in the last few years, the testimonials of many Zamboangueños suggest that there might be no room for optimism. ZC has been facing a multiglossic situation for decades and there are evidences that it is no longer the preferred language for socialisation among the younger generations, which raises some concerns regarding its future (Himoro, 2019). Besides (see Table 1), although the absolute number of native ZC speakers is still increasing, the census figures show that this increase is not proportional to the population growth, since the ratio of native ZC speakers in the Zamboanga City population is continuously decreasing. The on-going standardisation process is thus vital to enable the language to be used efficiently in higher written registers and counterbalance the hegemony of the official languages, namely English and Tagalog.

Year	%	No. of speakers	Population
1970	58.33%	116,611	199,901
1980	53.15%	182,701	343,722
1990	48.71%	215,490	442,345
2000	46.57%	280,252	601,794
2010	43.39%	350,240	807,129

Table 1: Ratio of native ZC speakers in the population of Zamboanga City according to 1970, 1980, 1990, 2000 and 2010 CPH. (National Census and Statistics Office (NCSO), 1974, 1983; National Statistics Office (NSO), 1992, 2003b, 2014b)

The approval of the orthography, nevertheless, has raised some other issues. The first of them regards its applicability: the main principle of the orthography is that words should be written according to their original form in the language of origin (DepEd Zamboanga City Division, 2016), which means that the spelling deemed as correct does not always reflect the phonetic evolution of the language, requiring prior study and memorisation work from the writers. The second one is basically technical: the gap between the different spelling systems currently in use in ZC and the orthography is often so big that the current *de facto* spell checking technologies are unable to yield acceptable results.

This research aims at addressing these issues and contributing to the standardisation of ZC by proposing i) a spelling error taxonomy for ZC formalised as an ontology and ii) an adaptive spell checking approach using Character-Based Statistical Machine Translation to correct spelling errors in ZC. We have also implemented a *hunspell* spell checker to be used as a baseline to evaluate our approach. We argue that this solution presents the following advantages over *hunspell*: (i) it can correct previously unseen words

by recognising some patterns, and (ii) it has a better performance when not dealing with simple spelling errors, but rather with different writing systems, as is the case of ZC.

2. The spelling errors ontology

2.1. Data

In order to study the spelling mistakes (i.e., errors) made by ZC speakers, we have built a corpus of written ZC, the Contemporary Written Zamboangueño Chabacano Corpus (CWZCC), containing 8,038,200 words and 9 text genres (namely Educational Texts, Fiction, Poems, Songs, News, Religion, Self-help, Internet and Misc) in two data formats: NLP Interchange Format (NIF) and Text Encoding Initiative XML (TEI-XML). ZC is mostly written in informal contexts (Himoro, 2019) and often a word is spelled in different ways even in the same text. In order to sufficiently capture this richness and/or variability, priority has been given to online sources, comprising about 70% of the corpus. Due to copyright and privacy restrictions, the corpus cannot be openly available to the public.

2.2. The error typology

In order to categorise the different spelling errors found, we applied an iterative process on samples of the CWZCC. For each iteration, we attempted to identify and classify the corresponding spelling errors. Whenever an error could not be placed under the existing categories, a new category was created. In this process, some categories were also divided into finer-grained ones or merged down. The spelling error categories thus identified are listed below (in alphabetical order). A code consisting of up to 3 characters has been assigned to each error for annotation purposes.

- Abbreviations (ABR)
- Cross-Linguistic Cognate Interference (COG)
- Euphemisms (EPH)
- Eye Dialect (ED)
- Inanities (INN)
- Insertion (INS)
- Lack of Capitalization (OC)
- Miscapitalization (XC)
- Misuse of the Apostrophe (XA)
- Misuse of the Hyphen (XH)
- Misuse of Spaces (XS)
- Omission (OMS)
- Omission: Apostrophe (OA)
- Omission: Hyphen (OH)
- Omission: Space (OS)
- Phoneme-Grapheme Mismatch (XPG)
- Repetitions (REP)
- Substitution (SUB)
- Transposition (TRS)
- Use of Diacritics (XD)
- Use of Homomorph Glyphs (HMM)
- Use of Homophone Graphemes (HOM)
- Use of Impossible Graphemes (XIG)
- Use of Inverted Punctuation Marks (XP)

2.3. The spelling error ontology assumptions, motivations and components

After identifying the spelling error categories, we proceeded to create a taxonomy that would allow us to unequivocally classify spelling errors in ZC. It is fundamentally based on the intentionality (fig. 1) and/or the cause of the errors and structured as an unbalanced hierarchy. In short, it grows in complexity as one moves deeper down in the tree, and less specific as one approaches the root node. This structure has been adopted for four reasons: 1) it leaves room to flexibility in the annotation, as cases in which classification might be problematic or ambiguous can be addressed by applying a less refined classification (using the category nodes located in the upper level in the hierarchy); 2) the proposed classification criteria can be more easily reused/replicated for languages in a similar situation; 3) a simpler and more general classification could be useful if it is decided to automatically annotate errors in the future, as it might be extremely difficult for machines to identify some of them; 4) when using NIF, it is highly recommended (at least, to full-comply with the recommendations for the development of linguistic linked data) to express the tagsets used in the corpus annotations with reference to one or more ontologies or other linked dataset. For more details about the corpus annotation, see section 4.
Despite the redundancy, for the sake of readability and clarity we have included images of each branch of the taxonomy. For the full scheme, see figure 14. The resulting scheme was later formalised as an OWL (Web Ontology Language) ontology [1].

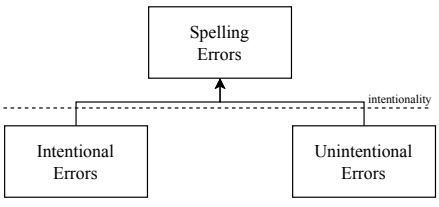

Figure 1: Top level concepts of the ZC spelling errors ontology.

1. **Intentional Errors**
   We classify under "Intentional Errors" (see fig. 2) a wide range of phenomena, including textisms (typical language found in text messages and internet lingo) and euphemisms. Textisms may be not strictly considered

[1] Available at https://research.chavacano.org/cwzcc.owl.

spelling errors, provided that (in most cases) users consciously decide whether to include a textism or not as a way to increase expressiveness or shorten messages. For this reason, textisms can be seen as a deviation from the correct spelling that, mostly, is not appropriate in formal writing, and this is why they have been included in this taxonomy.

1.1. **Abbreviation (ABR):** Error resulting from the omission of letters or the use of homophones letters and/or numbers to replace syllables.
∗kme → kame
∗solo2 → solo-solo
∗d2u → de tuyo
∗cge → segui

1.2. **Eye Dialect (ED):** Nonstandard written representation of a specific phonetic realization.
∗nema → no hay mas
∗pehcaw → pescao

1.3. **Inanities (INN):** Combinations of several words or nonsensical intentional transmogrification of words (Craig, 2003).
∗kers → quiere
∗dormz → dormi

1.4. **Repetition (REP):** Form of expression that indicates emotions such as astonishment, anger or excitement.
∗porkeeee → por que

1.5. **Use of Homomorph Glyphs (HMM):** Replacement of alphabetic characters with a similar-looking number or symbol.
∗k0sa → cosa

1.6. **Euphemism (EPH):** Intentional modification of a taboo word in order to mask expletives or profanity. A word is classified as such when it results in a nonexistent word in the language.
∗pota → puta
∗kunyununani → coño vos nana

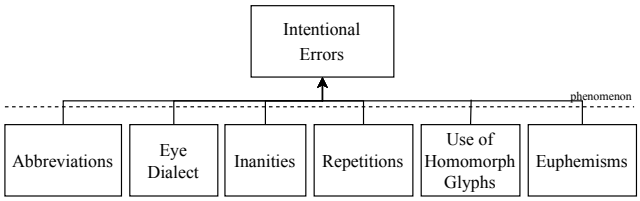

Figure 2: Subtypes of "Intentional Errors", second level of the ZC spelling errors ontology.

2. **Unintentional Errors**
Errors under this category occur essentially due to the speaker's lack of attention or knowledge. They are branched on awareness/randomness (see fig. 3).

2.1. **Random Errors**
These are errors whose cause is not likely to be other than a mere typo. Errors are only classified as such when other possible causes can be ruled out (shown in fig. 4).

2.1.1. **Insertion (INS):** Addition of a letter in a word without a plausible explanation.
∗karsa → casa

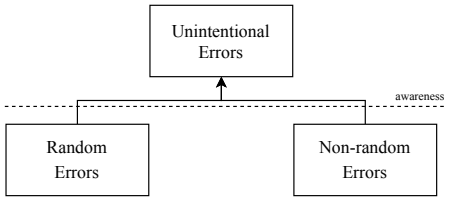

Figure 3: Subtypes of "Unintentional Errors", second level of the ZC spelling errors ontology.

2.1.2. **Omission (OMS):** Deletion of a letter from a word without a plausible explanation.
∗Chaacano → Chavacano

2.1.3. **Substitution (SUB):** Replacement of a letter in a word without a plausible explanation.
∗cpmpra → compra

2.1.4. **Transposition (TRS):** Swap of two letters in a word without a plausible explanation.
∗beuno → bueno

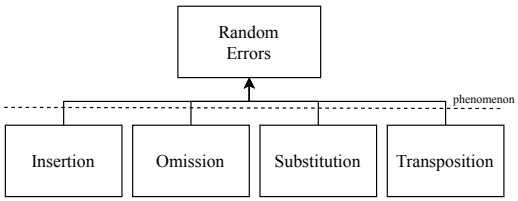

Figure 4: Subtypes of "Random Errors", third level of the ZC spelling errors ontology.

2.2. **Non-random Errors**
These are errors for which a likely cause is known. They are branched on the existence or not of practical rules that govern usage (see fig. 5).

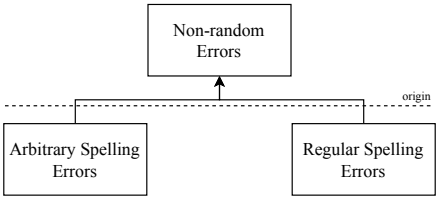

Figure 5: Subtypes of "Non-random Errors", third level of the ZC spelling errors ontology.

2.2.1. **Arbitrary Spelling[2] Errors**
Errors that affect aspects for which there are no practical rules that govern usage (see fig. 6). They are branched on phonetic plausibility, i.e., whether the resulting word can be pronounced as the word it actually represents or not.

2.2.1.1. **Phonogramical Errors**
Comprise the use of graphemes that represent the correct pronunciation of a word, but do not match those of the correct spelling (shown in fig. 7).

[2]The concepts of "Regular Spelling" and "Arbitrary Spelling" are based on Galí's typology (cited in Cristóbal Rojo, 1982).

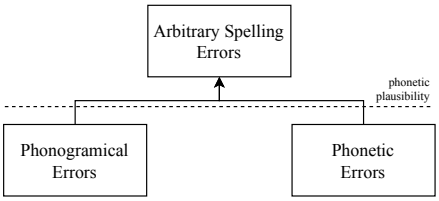

Figure 6: Subtypes of "Arbitrary Spelling Errors", fourth level of the ZC spelling errors ontology.

2.2.1.1.1. **Use of Homophone Graphemes (HOM):** Categorize the different uses of a grapheme that reflects a correct pronunciation of a word, but does not match the established correct spelling.
∗sapatos → zapatos
∗talya → talla

2.2.1.1.2. **Cross-Linguistic Cognate Interference (COG):** Partial or total spelling interference due to the existence of a cognate in one of the languages in contact with ZC.
∗attende → atende
∗technico → tecnico

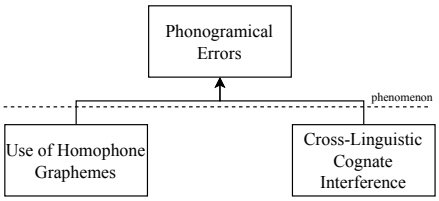

Figure 7: Subtypes of "Phonogramical Errors", fifth level of the ZC spelling errors ontology.

2.2.1.2. **Phonetic Errors**
Use of invalid graphemes in ZC or graphemes that do not allow the phonetic realization the speaker wants to represent (see fig. 8).

2.2.1.2.1. **Phoneme-Grapheme Mismatch (XPG):** Use of a grapheme that does not allow the phonetic realization the speaker wants to represent.
∗kunumon → kanamon

2.2.1.2.2. **Use of Impossible Graphemes (XIG):** Use of a grapheme not accepted in ZC.
∗qiere → quiere
∗itsura → hechura

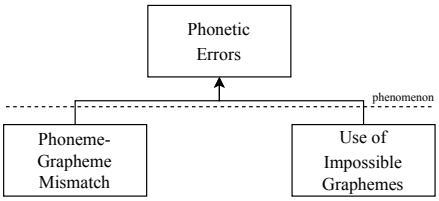

Figure 8: Subtypes of "Phonetic Errors", fifth level of the ZC spelling errors ontology.

2.2.2. **Regular Spelling[2] Errors**
Errors that affect regular aspects of the orthography (detailed in fig. 9).

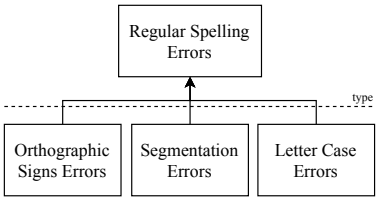

Figure 9: Subtypes of "Regular Spelling Errors", fourth level of the ZC spelling errors ontology.

2.2.2.1. **Orthographic Signs[3] Errors**
Auxiliary and punctuation mark errors (shown in fig. 10).

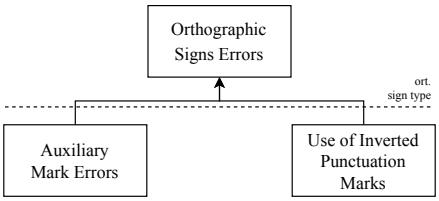

Figure 10: Subtypes of "Orthographic Signs Errors", fifth level of the ZC spelling errors ontology.

2.2.2.1.1. **Auxiliary Mark Errors**
Apostrophe, diacritics and hyphen errors (see fig. 11).

2.2.2.1.1.1. **Apostrophe Errors**
Incorrect use of the apostrophe.

2.2.2.1.1.1.1. **Apostrophe Omission (OA):** Missing apostrophe in a word that requires one.
∗tan → ta'n
∗unoy otro → uno'y otro

2.2.2.1.1.1.2. **Misuse of the Apostrophe (XA):** Use of the apostrophe in a word or between two words that do not require one.
∗sesenta'y nueve → sesenta y nueve

2.2.2.1.1.2. **Use of Diacritics (XD):** Use of diacritical marks due to Spanish or Tagalog influence.
∗mío → mio
∗olê → ole
∗galè → gale
∗vergüenza → verguenza

2.2.2.1.1.3. **Hyphen Errors**
Incorrect use of the hyphen.

2.2.2.1.1.3.1. **Hyphen Omission (OH):** Missing hyphen in a word or between two words that require one.
∗kosa kosa → cosa-cosa

2.2.2.1.1.3.2. **Misuse of the Hyphen (XH):** Use of the hyphen in a word or between two words that

[3]The concept of Orthographic Signs follows the Real Academia Española and Asociación de Academias de la Lengua Española (2005) definitions for Spanish. Thus, comma, quotes, square brackets, two points, question and exclamation marks, parenthesis, point, ellipsis, semicolon and dash are classified as Punctuation Marks. Apostrophe, asterisk, slash, diaeresis, hyphen, curly brackets, paragraph and acute accent, in turn, are classified as Auxiliary Signs.

do not require one.
∗alas-8 → a las 8
∗man-viaje → man viaje

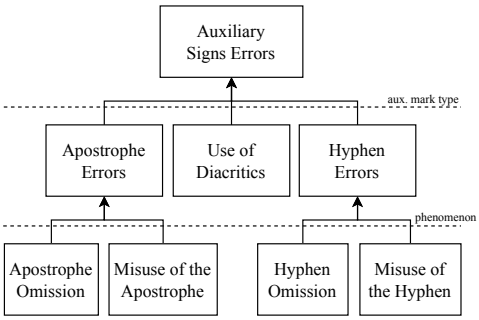

Figure 11: Subtypes of "Auxiliary Sign Errors", sixth level of the ZC spelling errors ontology.

2.2.2.1.2. **Use of Inverted Punctuation Marks (XP):** Use of Spanish inverted punctuation marks.
∗¿cosa? → cosa?
∗¡gracias! → gracias!

2.2.2.2. **Segmentation Errors**
Incorrect use of whitespaces (detailed in fig. 12).

2.2.2.2.1. **Space Omission (OS):** Missing space between words.
∗manmirahan → man mirahan
∗kunambre → con hambre
∗yalang → ya lang

2.2.2.2.2. **Misuse of Spaces (XS):** Use of spaces in a word that should not be segmented.
∗o hala → ojala

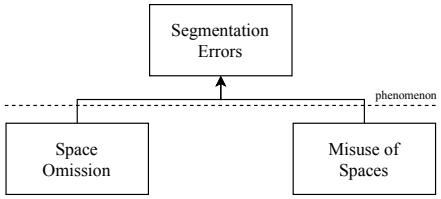

Figure 12: Subtypes of "Segmentation Errors", fifth level of the ZC spelling errors ontology.

2.2.2.3. **Letter Case Errors**
Incorrect use or lack of capitalisation (see fig. 13).

2.2.2.3.1. **Lack of Capitalisation (OC):** Missing capitalisation in a word that should be capitalised.
∗filipino → Filipino
∗pedro → Pedro

2.2.2.3.2. **Miscapitalisation (XC):** Capitalisation in a word that should not be capitalised.
Onde ∗Ustedes ta queda? →
Onde ustedes ta queda?

3. Spell checking tools

3.1. Character-based Statistical Machine Translation

Our approach consists of applying a Character-Based Statistical Machine Translation (SMT) model to detect and correct spelling errors. Unlike traditional Phrase-based MT (fig. 15), which translates sentences as sequences of words, Character-Based MT translates sequences of characters (fig. 16). The vocabulary is much smaller than in a phrase-based model, consisting of all letters (upper- and lower-case), numbers and symbols. There are, hence, no unknown words.

Basically, the model would, given an input word, apply the most probable transformations to it and output the transformed word in case it is incorrectly spelt, or keep it as-is in case it contains no spelling errors. The model was trained with individual words containing # and $ delimiters in the beginning and at the end respectively. In short, it should be able to distinguish if a given transformation is to be performed in the beginning, in the middle or at the end of the word, or generalise it in case it takes place in any position. In the example shown in fig. 16, the model would map the sequence of characters "ch" to "ti" in the beginning of words and "gge" to "gue" at the end of words. This setup enables the model to identify patterns and apply them to unseen words as well.

Figure 13: Subtypes of "Letter Case Errors", fifth level of the ZC spelling errors ontology.

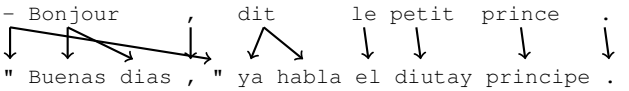

Figure 15: Example of lexical alignment in a traditional SMT system. (Sentences source: Saint-Exupéry [1943, p. 77], Saint-Exupéry [2018, p. 66]).

Figure 16: Two examples of lexical alignment in a Character-Based SMT system: each character is regarded as a word and multiwords are mapped to form patterns. # and $ characters are used in order to allow detecting patterns in the beginning and at the end of the word.

3.2. Related works

Character-Based SMT has been previously applied in different contexts, such as machine translation between related languages (Nakov & Tiedemann, 2012; Tiedemann, 2009; Vilar, Peter, & Ney, 2007), cognate production (Beinborn, Zesch, & Gurevych, 2013), historical (Korchagina, 2017; Schneider, Pettersson, & Percillier, 2017)

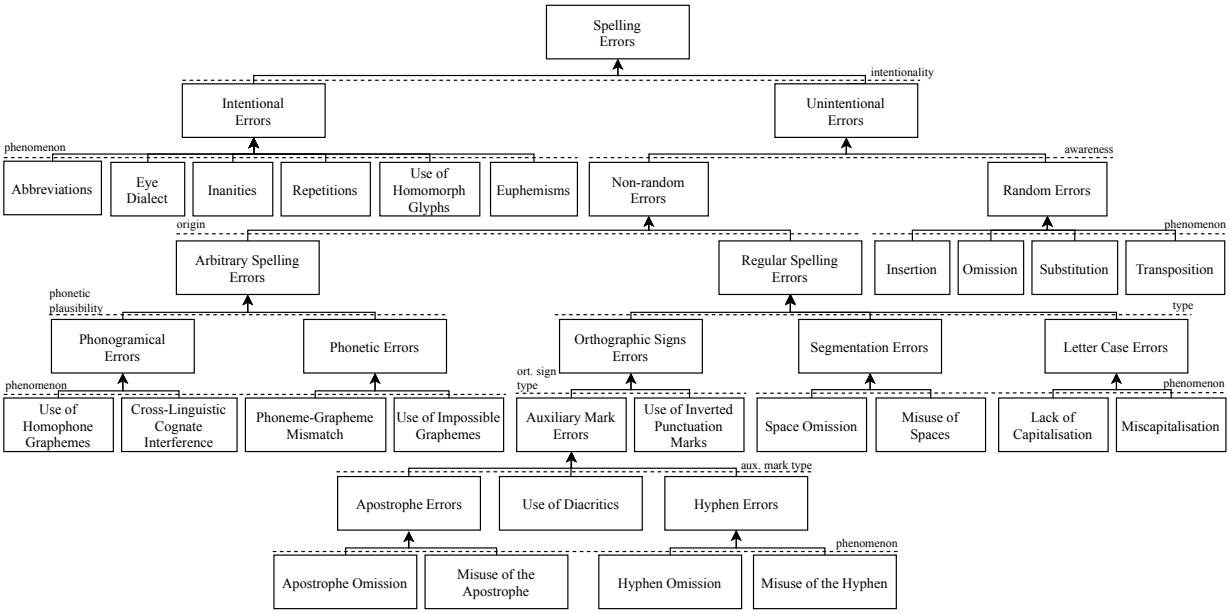

Figure 14: Complete taxonomy of the ZC spelling errors ontology

or dialectal (Scherrer, Samardžić, & Glaser, 2016) texts normalisation, transliterations (Karimi, 2008; Tiedemann & Nabende, 2009) and search queries spelling correction (Hasan, Heger, & Mansour, 2015).

3.3. Setup

For this experiment, we have used the Moses SMT engine (Koehn et al., 2007) with a SRILM 4-grams language model (n=4 empirically determined in our experiments) and GIZA++ as a word alignment tool.

3.4. Training data

In order to train a spell checking tool for ZC, a total of about 100,000 words has been extracted from CWZCC. Spelling errors have been manually corrected and assigned to at least one of the categories of the error taxonomy. Given most of the time multiple errors occur in a single word, misspelt words may be labelled with multiples categories.

3.5. Limitations

Although we have corrected spelling errors that affect more than one token, these entries have been removed from the training data set. In other words, the model has been trained exclusively on unigrams. Exception is only made to words that were mistakenly written as a single word but should be written separated by spaces and in the source comprise a single token. We also have refrained from correcting punctuation mark errors other than inverted punctuation marks, which are not part of ZC. We do not correct barbarisms, interference and hypercorrection errors unless they are spelling errors.

3.6. *hunspell* (baseline)

hunspell is nowadays one of the most popular open source spell checkers (Németh, 2019), and hence, used as a baseline in this research. Given that one of the prerequisites to build a *hunspell* spell checker is a wordlist, we have extracted words from 5 lexicographic works for ZC and manually corrected their spelling to comply with the official orthography. The final wordlist consists of 13,167 words.

In the affix file, we have included the most frequent suffixes and prefixes with a relatively regular behaviour. The implemented prefixes were: **aka-**-, **ika-**-, **ma-**, **maka-**, **pagka-**, **ka-**, **pinaka-**. As for the suffixes, we have implemented: **-han**, infinitive (**-r** after prepositions, as frequently observed in the speech of some elderly), gerund (**-ando**, **-iendo**) and participle (**-ao**, **-ido**) verb endings and nominalisers (**-ada**, **-ida**). No replacement or phonetic rules have been implemented.

4. Corpus annotation

Once the SMT model was trained, we have built a tool written in Python to automatically annotate our corpus. The tool takes CWZCC original files as input and makes XML-RPC requests to `mosesserver` using the trained model. Errors that have been previously annotated in the training phase receive their respective labels. Unannotated errors get the "UND" label (undefined).

In order to generate an annotated NIF version of the corpus, a modified version of the package pynif has been used. The output are 9 .ttl files, each corresponding to one text genre in the corpus, and linked to the spelling errors OWL ontology developed in this research. Misspelt tokens of every document in the genre are annotated with its candidate corrected form using the property `correction` and its corresponding classification according to the taxonomy using the property `classAnotation`. A minimal example can be found in Annex A.

In the TEI-XML version of the corpus, every genre consists of multiples files, each one corresponding to one document. For the sake of simplicity, a document structure complying with the TEI Minimal Header has been used. Misspelt tokens are annotated using the tag `error` and the multivalue attribute `@type`, which contains its corresponding classification according to the taxonomy, and the attribute

`@correction`, which contains the candidate corrected form. An example error-annotated document of the TEI-XML version can be found in Annex B.

5. Results

5.1. Spelling error statistics

Table 2 shows the number of occurrences for each category of spelling error. The big picture can be seen in table 3, which shows the number of occurrences by major error type.

Code	Occurrences
HOM	13,605
OA	2,933
OS	2,623
ED	1,909
ABR	1,610
OC	1,351
COG	1,088
XC	588
OMS	450
REP	389
XD	347
INS	307
XIG	234
XS	223
SUB	161
OH	143
TRS	104
XA	54
EPH	39
XH	38
INN	27
XP	21
XPG	10
HMM	4
Total	28,258

Table 2: Occurrences of each spelling error categories in the training data.

Error Type	Occurrences	%
Intentional Errors	3,978	14.08
Unintentional Errors	24,280	85.92
Random Errors	1,022	4.21 (3.62)
Non-random Errors	23,258	95.79 (82.30)
Arbitrary Sp. Err.	14,937	64.22 (52.86)
Regular Sp. Err.	8,321	35.78 (29.44)

Table 3: Occurrences of spelling errors by major error type. The percentages show the frequency of occurrence of each error type in the same level of the hierarchy, while the values given in parenthesis are its frequency over the total of spelling errors.

As shown in these tables, Unintentional Errors make up for the majority of the spelling errors found, from which most are Non-random Errors. The most frequent error is "Use of Homophone Graphemes" (HOM), an Arbitrary Spelling Error. This was expected, given the asymmetry between the large inventory of graphemes a ZC speaker has at his or her disposal to represent a limited number of phonemes. In a more fragmented way, follow various Regular Spelling Errors, which indicates that, once ZC speakers learn the rules of the new orthography, a substantial part of the spelling errors should become less frequent. "Eye Dialect Errors", classified under Intentional Errors, are also quite numerous, highlighting the distance between the written and the oral forms of some words and expressions. Finally, a relatively high occurrence of "Abbreviation Errors", a typical Intentional Error often found in informal writings, can be explained by the fact that much of the sample data comes from social networks.

5.2. Spell checkers

In order to evaluate both spell checkers, k-fold cross validation with k = 10 folds has been used, considering only the top 3 candidates. Tables 4 and 5 show different metrics for both *hunspell* and our approach using Character-based SMT. First, it can be noted that the values found for *hunspell* in all metrics were significantly lower than those found for the Character-Based SMT model. The precision values for *hunspell* show that it performed reasonably well at correcting spelling errors. However, the low recall value indicates that many of the words detected as spelling errors are actually not real errors. For all metrics, the Character-Based SMT model outperforms *hunspell* for the dataset used in this research.

Candidates considered	top 1 candidate	top 2 candidates	top 3 candidates
Accuracy	59,67%	61,62%	63,84%
Precision	77,94%	80,16%	82,84%
Recall	68,9%	68,9%	68,9%
F-measure	73,14%	74,1%	75,23%

Table 4: Accuracy, Precision, Recall and F-measure values for the *hunspell*-based spell checker (baseline) according to the number of candidates considered.

Candidates considered	top 1 candidate	top 2 candidates	top 3 candidates
Accuracy	88,26%	93,6%	94,9%
Precision	93,01%	96,38%	97,19%
Recall	92,22%	95,58%	96,41%
F-measure	92,61%	95,98%	96,8%

Table 5: Accuracy, Precision, Recall and F-measure values for the spell checker using Character-Based SMT model according to the number of candidates considered.

The main reason for the poorer performance of the *hunspell*-based spell checker is the high occurrence of code mixing and code-switching in ZC. When a word is not found in the wordlist, it is automatically marked as incorrect, as is the case of words from English, Tagalog, Ce-

buano and other languages. In that sense, the SMT model far outweighs *hunspell*, since it is capable of generating both ZC words and those of other languages, as long as they have appeared in the training data. It is also able to offer plausible corrections for unknown words, based on patterns found in the training data, although in some cases it might suggest invalid candidates or distort unknown words that need no correction due to an overgeneralisation of the rules learned in the training phase. The orthographic distance between the incorrect and the correct forms also plays a factor in the low performance of *hunspell*. Quite often the correct form does not make it to the first three candidates, or is not suggested at all even when considering a higher number of candidates. Although the reliance of *hunspell* on a wordlist might be seen as a weak point, given only words that appear in the wordlist can be suggested as correction candidates, it is at the same time its strong point, since it guarantees no invalid words will ever be suggested as corrections.

On the other hand, *hunspell* seems to perform better for random errors, that is, those in which a letter in the word is inserted, deleted, transposed or replaced. In these cases, the SMT model often ends up deforming the word and generating invalid candidates, unless similar cases have appeared in the training data. Another case in which *hunspell* could outperform the SMT model is with rarely used or register-specific words. In addition, *hunspell* is scalable and can produce a real-time response while text is still being inputted by the user. With Moses, however, obtaining a real-time response would not be feasible, forcing the user to input the whole text to be checked, and only after that check it using the spell checker. Finally, the main advantage of *hunspell* is its high degree of interoperability with other programs and applications, which eliminates the need for further development. Considering all this, the ideal situation would be to combine both approaches.

6. Conclusion

In this research, we have built a written Zamboanga Chabacano corpus and, by studying the spelling errors speakers make considering the orthography as the reference, we have developed a spelling error taxonomy formalised as an ontology. We trained a Character-Based Statistical Machine Translation model using manually corrected data from the corpus and showed that this approach outperforms *hunspell* for the case of Zamboanga Chabacano. However, there is still much work to be done.

As a follow-up of this study, we propose extracting phonetic and substitution rules from Moses' phrase tables and, after pruning and removing spurious data, introducing them in the *hunspell* affix file. It would be equally interesting to come up with ways of dealing with homophone words or errors that affect multiple tokens. A possible solution could be training a mixed unigram-bigram model or training two independent models and combining their outputs. Since throughout the course of this project we have not had access to the official Zamboanga Chabacano dictionary recently published by the local government, adjusting the forms used both in the wordlist and in the training data is an urgent task if we ever gain access to it. Other future tasks include semi-automatically correcting the rest of corpus in an incremental way in order to obtain more training data, expanding the wordlist with words from the training data, further expanding the corpus and trying to apply the same approach using a Neural Machine Translation model.

7. Acknowledgements

This work has been supported by the projects **CetrO+Spec** (Creation, Exploration and Transformation of Educational Object Repositories in Specialized Domains, ref. TIN2017-88092-R) and **SWITCHED-ON** (The empowerment of massive open social language learning through mobile technology: harnessing interactions, transcending boundaries, ref. FFI2016-80613-P), both of them funded by the Spanish Ministry of Economy and Competitiveness.

8. Bibliographical References

Beinborn, L., Zesch, T., and Gurevych, I. (2013). Cognate Production using Character-based Machine Translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, January.

Craig, D. (2003). Instant messaging: the language of youth literacy. In *The Boothe Prize Essays 2003*. Stanford University.

Cristóbal Rojo, M. (1982). Mi descubrimiento sobre Alexandre Galí. *Maina*, (5):56–59.

DepEd Zamboanga City Division. (2016). *Revised Zamboanga Chavacano Orthography (Guia para na Enseñanza de Chavacano)*. Zamboanga City Local Government, Zamboanga.

Government of the Philippines. (2011). DepEd develops learning supplements using mother-tongue, retrieved January 19, 2017, from https://www.officialgazette.gov.ph/2011/11/28/deped-develops-learning-supplements-using-mother-tongue/

Hasan, S., Heger, C., and Mansour, S. (2015). Spelling Correction of User Search Queries through Statistical Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*,

Himoro, M. Y. (2019). *Hacia un corrector ortográfico para la nueva ortografía del chabacano de Zamboanga*. Master's thesis, Universidad Nacional de Educación a Distancia, September.

Karimi, S. (2008). *Machine transliteration of proper names between English and Persian*. Ph.D. thesis RMIT University, Melbourne, January.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180, Prague, Czech Republic, June. Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P07-2045

Komisyon sa Wikang Filipino (KWF). (2015). Mapa ng mga Wika ng Filipinas, retrieved August 3, 2018, from http://kwf.gov.ph/mapa-ng-mga-wika-ng-filipinas/

Korchagina, N. (2017). Normalizing Medieval German Texts: from rules to deep learning. In *NoDaLiDa 2017 Workshop on Processing Historical Language*, Gothenburg, May.

Nakov, P., and Tiedemann, J. (2012). Combining Word-Level and Character-Level Models for Machine Translation Between Closely-Related Languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Vol. 2, pp. 301–305. Jeju Island, July.

National Census and Statistics Office (NCSO). (1974). *1970 Census of Population and Housing, Final Report - Vol. 1 - Zamboanga del Sur*, Manila, March. Retrieved from https://psa.gov.ph/content/census-population-and-housing-report

National Census and Statistics Office (NCSO). (1983). *1980 Census of Population and Housing, Volume 1, Final Report - Zamboanga del Sur*, Manila, Manila, Filipinas, May. Retrieved from https://psa.gov.ph/content/census-population-and-housing-report

National Statistics Office (NSO). (1992). *1990 Census of Population and Housing, Report No. 3 - 86 I - Socio-Economic and Demographic Characteristics*, Manila, June. Retrieved from https://psa.gov.ph/content/census-population-and-housing-report

National Statistics Office (NSO). (2003a). *2000 Census of Population and Housing, Report No. 2 Vol. 1 - Demographic and Housing Characteristics*, Manila, January. Retrieved from https://psa.gov.ph/content/census-population-and-housing-report

National Statistics Office (NSO). (2003b). *2000 Census of Population and Housing, Report No. 2 Vol. 1 - Demographic and Housing Characteristics - Zamboanga City*, Manila, January. Retrieved from https://psa.gov.ph/content/census-population-and-housing-report

National Statistics Office (NSO). (2014a). *2010 Census of Population and Housing, Report No. 2B - Population and Household Characteristics (Sample Variables)*, Manila, February. Retrieved from https://psa.gov.ph/content/census-population-and-housing-report

National Statistics Office (NSO). (2014b). *2010 Census of Population and Housing, Report No. 2B - Population and Household Characteristics (Sample Variables) - Zamboanga City*, Manila, February. Retrieved from https://psa.gov.ph/content/census-population-and-housing-report

Németh, L. (2019, July 29). (Version 0.60.7), July. Retrieved from http://aspell.net/

Philippine Statistics Authority (PSA). (2014). "Statistical Tables on Sample Variables from the Results of 2010 Census of Population and Housing" in Census of Population and Housing, July. Retrieved from https://psa.gov.ph/population-and-housing/statistical-tables/2010

Real Academia Española, and Asociación de Academias de la Lengua Española. (2005). *Diccionario panhispánico de dudas*. Real Academia Española.

Saint-Exupéry, A. (1943). *Le Petit Prince*. Ebooks libres et gratuits. Retrieved from https://www.cmls.polytechnique.fr/perso/tringali/documents/st_exupery_le_petit_prince.pdf

Saint-Exupéry, A. (2018). *El Diutay Principe (trans. Jerome Herrera)*. Jerome Herrera.

Scherrer, Y., Samardžić, T., and Glaser, E. (2016). Normalizing orthographic and dialectal variants in the ArchiMob corpus of spoken Swiss German. ID: unige:90850, retrieved from https://archive-ouverte.unige.ch/unige:90850

Schneider, G., Pettersson, E., and Percillier, M. (2017). Comparing Rule-based and SMT-based Spelling Normalisation for English Historical Texts. In *NoDaLiDa 2017 Workshop on Processing Historical Language*, May.

Tiedemann, J. (2009). Character-based PSMT for Closely Related Languages. In *Proceedings of the 13th Annual Conference of the EAMT*, pp. 12–19. Barcelona, May.

Tiedemann, J., and Nabende, P. (2009). Translating Transliterations. *International Journal of Computing and ICT Research*, *3*:33–41.

Vilar, D., Peter, J.-T., and Ney, H. (2007). Can We Translate Letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 33–39, Prague, Czech Republic, Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1626355.1626360

Annex A:
Minimal sample of error annotation, extracted from the NIF version of the corpus

This is an example of the first lines of a NIF file of the corpus (one per genre).

```
@prefix cwzcc: <http://research.chavacano.org/cwzcc.owl#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```

The next block represents the existing documents in the genre. Every document is univocally identified by a URL.

```
<http://research.chavacano.org/cwzcc> a nif:ContextCollection ;
   nif:hasContext
      <http://research.chavacano.org/cwzcc/news/1-1-bilyones-de-pesos-budget-na-security-
         ya-hace-bandera-el-alcalde-climaco>,
      <http://research.chavacano.org/cwzcc/news/1-2-kilo-shabu-confiscao>,
      ...
      <http://research.chavacano.org/cwzcc/news/zcwd-ta-asegura-cay-acava-el-pipelaying-
         na-fin-de-este-mez> ;
   dcterms:conformsTo
      <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/2.1> .
```

Every document in the genre is then declared as follows:

```
<http://research.chavacano.org/cwzcc/social/twitter/14> a nif:Context,
      nif:OffsetBasedString ;
   nif:beginIndex "0"^^xsd:nonNegativeInteger ;
   nif:endIndex "140"^^xsd:nonNegativeInteger ;
   nif:isString "pirmi man iyo ta sinti dwele.. nusabe yo porke pirmi ansina.. kwando iyo
       keda alegre? kwando gaha pasa el diya o mes na hinde iyo triste? =(" ;
   dcterms:date "2009-08-27 00:30"^^xsd:string .
```

Finally, misspelt tokens of every document are annotated with its correct form using the property `correction` and its corresponding classification according to the taxonomy using the property `classAnotation`.

```
<http://research.chavacano.org/cwzcc/social/twitter/14#offset_0_5>
   a nif:OffsetBasedString,
     nif:Phrase ;
   nif:anchorOf "pirmi" ;
   nif:beginIndex "0"^^xsd:nonNegativeInteger ;
   nif:endIndex "5"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:COG ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "firme"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_10_13>
   a nif:OffsetBasedString,
     nif:Phrase ;
   nif:anchorOf "iyo" ;
   nif:beginIndex "10"^^xsd:nonNegativeInteger ;
   nif:endIndex "13"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
```

```
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "yo"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_17_22>
   a nif:OffsetBasedString,
     nif:Phrase ;
   nif:anchorOf "sinti" ;
   nif:beginIndex "17"^^xsd:nonNegativeInteger ;
   nif:endIndex "22"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "senti"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_23_28>
   a nif:OffsetBasedString,
     nif:Phrase ;
   nif:anchorOf "dwele" ;
   nif:beginIndex "23"^^xsd:nonNegativeInteger ;
   nif:endIndex "28"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "duele"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_31_37>
   a nif:OffsetBasedString,
     nif:Phrase ;
   nif:anchorOf "nusabe" ;
   nif:beginIndex "31"^^xsd:nonNegativeInteger ;
   nif:endIndex "37"^^xsd:nonNegativeInteger ;
   nif:classAnotation [cwzcc:HOM, cwzcc:OS] ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "no sabe"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_41_46>
   a nif:OffsetBasedString,
     nif:Phrase ;
   nif:anchorOf "porke" ;
   nif:beginIndex "41"^^xsd:nonNegativeInteger ;
   nif:endIndex "46"^^xsd:nonNegativeInteger ;
   nif:classAnotation [cwzcc:HOM, cwzcc:OS] ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "por que"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_47_52>
   a nif:OffsetBasedString,
     nif:Phrase ;
   nif:anchorOf "pirmi" ;
   nif:beginIndex "47"^^xsd:nonNegativeInteger ;
   nif:endIndex "52"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:COG ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "firme"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_62_68>
   a nif:OffsetBasedString,
     nif:Phrase ;
   nif:anchorOf "kwando" ;
   nif:beginIndex "62"^^xsd:nonNegativeInteger ;
   nif:endIndex "68"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "cuando"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_69_72>
   a nif:OffsetBasedString,
```

```
      nif:Phrase ;
   nif:anchorOf "iyo" ;
   nif:beginIndex "69"^^xsd:nonNegativeInteger ;
   nif:endIndex "72"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "yo"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_73_77>
   a nif:OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "keda" ;
   nif:beginIndex "73"^^xsd:nonNegativeInteger ;
   nif:endIndex "77"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "queda"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_86_92>
   a nif:OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "kwando" ;
   nif:beginIndex "86"^^xsd:nonNegativeInteger ;
   nif:endIndex "92"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "cuando"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_106_110>
   a nif:OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "diya" ;
   nif:beginIndex "106"^^xsd:nonNegativeInteger ;
   nif:endIndex "110"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "dia"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_120_125>
   a nif:OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "hinde" ;
   nif:beginIndex "120"^^xsd:nonNegativeInteger ;
   nif:endIndex "125"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "hende"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_126_129>
   a nif:OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "iyo" ;
   nif:beginIndex "126"^^xsd:nonNegativeInteger ;
   nif:endIndex "129"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14> ;
   cwzcc:correction "yo"^^xsd:string .
```

Annex B:
Example of an error-annotated TEI-XML document, extracted from the TEI-XML version of the corpus, corresponding to the errors annotated in Annex A

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>twitter14</title>
        <author />
      </titleStmt>
      <editionStmt />
      <publicationStmt>
        <p>Published as a part of the <title ref="#HC_XML">Helsinki Corpus TEI XML
            Edition</title>.</p>
        <date when="2009-08-27 00:30"/>
      </publicationStmt>
    </fileDesc>
    <revisionDesc>
      <listChange>
        <change>
          <name />
          <date />
        </change>
      </listChange>
    </revisionDesc>
  </teiHeader>
  <text>
    <body>
      <p><error type="COG" correction="firme">pirmi</error> man <error type="HOM"
          correction="yo">iyo</error> ta <error type="HOM" correction="senti">sinti</error
          > <error type="HOM" correction="duele">dwele</error>.. <error type="HOM,OS"
          correction="no sabe">nusabe</error> yo <error type="HOM,OS" correction="por que"
          >porke</error> <error type="COG" correction="firme">pirmi</error> ansina.. <
          error type="HOM" correction="cuando">kwando</error> <error type="HOM" correction
          ="yo">iyo</error> <error type="HOM" correction="queda">keda</error> alegre? <
          error type="HOM" correction="cuando">kwando</error> gaha pasa el <error type="
          HOM" correction="dia">diya</error> o mes na <error type="HOM" correction="hende"
          >hinde</error> <error type="HOM" correction="yo">iyo</error> triste? =(</p>
    </body>
  </text>
</TEI>
```