

# Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalisation

Alejandro Mosquera, Elena Lloret, Paloma Moreda

University Of Alicante  
DLSI. Ap.de Correos 99. E-03080 Alicante, Spain  
amosquera@dlsi.ua.es, elloret@dlsi.ua.es, moreda@dlsi.ua.es

## Abstract

The Web 2.0, through its different platforms, such as blogs, social networks, microblogs, or forums allows users to freely write content on the Internet, with the purpose to provide, share and use information. However, the non-standard features of the language used in Web 2.0 publications can make social media content less accessible than traditional texts. For this reason we propose TENOR, a multilingual lexical approach for normalising Web 2.0 texts. Given a noisy sentence either in Spanish or English, our aim is to transform it into its canonical form, so that it can be easily understood by any person or text simplification tools. Our experimental results show that TENOR is an adequate tool for this task, facilitating text simplification with current NLP tools when required and also making Web 2.0 texts more accessible to people unfamiliar with these text types.

**Keywords:** Accessibility, Normalisation, Web 2.0

## 1. Introduction

The Web 2.0, through its different platforms, such as blogs, social networks, microblogs, or forums allows users to freely write content on the Internet, with the purpose to provide, share and use information. It is known that this type of platforms are among the top visited websites<sup>1</sup>, and their interest is growing more and more.

However, despite of the great potential of this user-generated content, it has several well-known drawbacks, concerning what is communicated and how it is communicated. On the one hand, the information users provide has not always the same level of reliability, and therefore wrong or inaccurate information can be considered as correct one (Scanfeld et al., 2010), (Mendoza et al., 2010). On the other hand, the Internet, and in particular, the Web 2.0, has an informal nature, since there is not any restriction regarding the language employed for posting on-line information. To name just a few: i) the use of emoticons (e.g., :-P); ii) non-standard abbreviations (e.g., LOL – *laugh out loud*) and contractions (e.g., abt – *about*); iii) frequent typos and spelling errors (e.g., *lasi*, instead of *lazy*); and iv) a lot of use of interjections and letter-repetitions (e.g., *yeeah-hhhhhh!*).

These non-standard features can make Web 2.0 publications less accessible than traditional texts to people unfamiliar with this type of lexical variants or people with disabilities. To date and to our knowledge, studies on text accessibility focus on simplification strategies. For this reason, performing a normalisation process is a step prior to simplification for non-accessible Web 2.0 texts.

Therefore, the objective of this paper is to suggest TENOR, a lexical approach for normalising Web 2.0 texts. Given a noisy sentence either in Spanish or English, our aim is to transform it into its canonical form, so that it can be easily understood by any person or text simplification tools. By achieving this goal, texts could be transcribed using

standard and common language, making them easier and more comprehensible, and thus facilitating straightforward the reading comprehension process for people with difficulties, as well as the use of existing automatic tools for carrying out other tasks, such as text simplification or summarisation.

This article is organised as follows. In Section 2, the state of the art is reviewed, discussing existing research works dealing with text simplification and normalisation, and stressing the differences of our approach with respect to them. Further on, Section 3 describes our normalisation approach for very informal texts. Next, in Section 4 we described the evaluation conducted, together with a in-depth discussion of the results obtained, and finally Section 5 concludes this paper and outlines future work.

## 2. Related Work

In the recent years, making information more accessible to everybody is a relevant issue which is gaining a lot of attention among the research community. One of the research areas devoted to this purpose is Text Simplification whose aim is to rewrite the information into a simpler way in order to help users to comprehend the information that, if left unedited, would be too complex to understand. To this end, the types of simplification include: i) lexical, which substitutes non-frequent words to more common ones (Biran et al., 2011); ii) syntactic, which splits difficult and large sentences into simpler ones (Evans, 2011); and iii) semantic, which attempts to provide definitions for difficult expressions and/or non-literal meaning (Barnden, 2008). Initiatives such as Simple Wikipedia<sup>2</sup>, Noticias Fácil<sup>3</sup> as well as several past and on-going projects, as for instance, Skill-Sum (Williams and Reiter, 2008), Simplext (Saggion et al., 2011), or FIRST<sup>4</sup>, constitute good contexts for mak-

<sup>1</sup><http://www.alexa.com/topsites>

<sup>2</sup>[http://simple.wikipedia.org/wiki/Main\\_Page](http://simple.wikipedia.org/wiki/Main_Page)

<sup>3</sup><http://www.noticiasfacil.es/ES/Paginas/index.aspx>

<sup>4</sup><http://www.first-asd.eu/>

ing progress within this area, thus being beneficial for individuals with low literacy (Candido et al., 2009), physical and cognitive disabilities (Daelemans et al., 2004), (Huenerfauth et al., 2009), or even language learners (Petersen and Ostendorf, 2007).

However, as these systems are designed to work with standard texts, the special features of the language used in the Web 2.0 can difficult their processing.

Furthermore, another subfield of Natural Language Processing (NLP) deals with Text Normalisation of user-generated content.

The process of text normalisation basically cleans an input word or sentence by transforming all non-standard lexical or syntactic variations into their canonical forms. From the existing literature, we have identified three major trends to tackle this task. The first one relies on machine translation techniques (Aw et al., 2006), (López et al., 2010) the second focuses on orthographic correction approaches (Liu et al., 2011), and the third one takes as a basis a combination of lexical and phonetic edit distances (Han and Baldwin, 2011), (Gouws et al., 2011). Among them, we would like to outline the research works proposed in (Han and Baldwin, 2011) and (Liu et al., 2011). In the former, supervised classification techniques are employed for identifying ill-formed words, which are then normalised by extracting the best candidate among several ones, using a set of rules. In the latter, a letter transformation approach is proposed through the use of the noisy channel model (Shannon, 1948).

To the best of our knowledge, none of the previous works have used a multilingual strategy, thus being restricted to the English language only. Therefore, this paper proposes the use of TENOR, a multilingual normalisation tool for the Web 2.0 with the purpose of obtaining the canonical form of a text, so it can be more accessible to more people and for current NLP simplification tools.

In the next section, our approach will be explained in detail.

### 3. TENOR, Text Normalisation Approach

In this section we explain TENOR, our text normalisation approach based on a combination of lexical and phonetic edit distances for short English and Spanish texts belonging to the Web 2.0.

Our normalisation process comprises two steps: First, it uses a classification method to detect non-standard lexical variants or words out of vocabulary. Second, the selected words in the previous step are replaced to their original standard form. Each of this stages are going to be explained in more detail.

#### 3.1. Out-of-Vocabulary detection

In this section we refer to words outside the vocabulary as those that are not part of standard English or Spanish vocabulary and need to be standardised. However, the detection of such words is not a trivial task: The presence of proper names, cities, neologisms and acronyms, as well as the richness of the language makes it difficult to know when a word belongs to the language or otherwise is a lexical variant (see Table 1).

	OOV word	Canonic word
a)	sucess	success
b)	rite	right
c)	playin	playing
d)	emocion	emoción
e)	mimir	dormir
f)	separa2	separados

Table 1: Out of vocabulary and canonic pairs examples from Web 2.0 texts. Examples from *a* to *c* correspond to English and the ones from *d* to *f* to Spanish.

In TENOR, OOV words are detected with a dictionary lookup. In order to do this, we use custom-made lexicons built over the expanded English and Spanish Aspell<sup>5</sup> dictionaries. These are augmented with domain-specific knowledge such as the Spell Checking Oriented Word Lists (SCOWL)<sup>6</sup> for English, and country names, cities, acronyms and common proper names<sup>7</sup> for Spanish. Heuristics based on capitalisation of words are employed to identify named entities and acronyms. Likewise, some special Twitter tags are used to perform a slight syntactic disambiguation, such as: @(User Name) # (Tag), RT (Retweet) and TT (Trending Topic), thus avoiding the processing of such elements.

#### 3.2. Substitution of Lexical Variants

This section discusses the different steps carried out to replace the words classified as OOV with their normalised form. In order to do this, several substages are proposed. First, in Section 3.2.1 the filtering techniques employed to “clean” texts are introduced. In Section 3.2.2 we detail the process of replacing common word transformations. Then, in Section 3.2.3 the use of phonetic indexing in order to obtain lists of words with equivalent pronunciations by building a phone lattice is described. Subsequently, in Section 3.2.4 we explain how this lattice is used in order to identify possible candidates to replace the non-normative lexical variants. Finally, in Section 3.2.5 we show how the use of language models can help to select the most appropriate canonical word from the list of phonetic candidates.

##### 3.2.1. Filtering

First, all non-printable characters and non-standard punctuations with the exception of emoticons are eliminated using regular expressions. While these may be beyond the scope of the study and therefore not to be considered lexical variants, their filtering would negatively impact another NLP tools such as opinion mining or sentiment analysis.

##### 3.2.2. Common Word Transformations

The second step of the analysis is to identify common word transformations such as abbreviations and transliterations, which are replaced by their equivalent standard form: i) Word-lengthening compression (see Table 3, example c) is

<sup>5</sup><http://aspell.net>

<sup>6</sup><http://wordlist.sourceforge.net/>

<sup>7</sup><http://es.wikipedia.org>

performed by applying heuristic rules to reduce the repetition of vowels or consonants within a word (*nooo! - no!*, *gooooooolll - gol!*); ii) There are numbers whose pronunciation is often used to shorten the length of the message (*ning1 - ninguno*) or combination of letters and (*h0us3 - house*). In these cases they were replaced by following a transliteration conversion table. In Table 2, each number is assigned its most frequent meanings when it appears as a part of a word; iii) Emoticon translation (see Table 3, example b) was made by grouping smileys into two categories (happy, sad), thus being replaced by their textual equivalent using simple heuristic rules based on regular expressions; iv) Simple case restoration techniques were applied to wrong-cased words (*GrEaT - great*).

Nº	English	Spanish
0	0, zero, o	0, cero, o
1	1, one	1, uno
2	2, two, too	2, dos
3	3, three, e	3, tres, e
4	4, for, a	4, cuatro, a
5	5, five, s	5, cinco, s
6	6, six, g	6, seis, g
7	7, seven, t	7, siete, t
8	8, eight	8, ocho
9	9, nine, g	9, nueve, g

Table 2: Common numeric transliterations found in Web 2.0 English and Spanish texts.

### 3.2.3. Phonetic Indexing

The aim of this stage is to obtain a list of candidate terms for each OOV words detected in previous stages. In order to do this, TENOR obtains lists of words with equivalent pronunciations using phonetic indexing techniques to build a phone lattice. OOV words are matched against this phone lattice with the metaphone algorithm (Philips, 2000) to obtain such list of substitution candidates. The metaphone algorithm allows to represent the pronunciation of a word using a set of rules. In particular the double-metaphone reference implementation for English and an adaptation of the metaphone for the Spanish language<sup>8</sup>. For example, the Spanish metaphone (*JNTS*) can index the words *gentes*, *jinetas*, *jinetes*, *juanetes*, *juntas*, *juntos* between others and the English metaphone (*PRXS*) can index the words *purses*, *prices*, *precise*, *praises* among others.

Moreover, there are acronyms and abbreviated forms that can not be detected properly with phonetic indexing techniques (*lol - laugh out loud*). For this reason, TENOR uses an exception dictionary manually built upon an equivalence table with 46 of the most common Spanish abbreviations (*qta - qué tal*), (*xfa - por favor*) and 196 English Internet abbreviations and slang words<sup>9</sup> that need special treatment because their low similarity with their equivalent standard form (*gotta - going to*), (*omg - oh my god*).

### 3.2.4. Lexical Similarity

Once the possible candidates associated to a OOV word are obtained, the lexical similarity between each candidate and the OOV word is computed. For this, we use the Gestalt pattern matching algorithm (Ratcliff and Metzner, 1988). This algorithm provides a string similarity score based on the maximum common subsequence principle between 0 and 100, where 0 is minimum similarity and 100 is maximum similarity. This score is calculated between the OOV word and its candidate list, empirically discarding candidates with similarity values lower than 60.

### 3.2.5. Candidate Selection

In order to obtain the final substitution candidate when there are more than one candidate word with the same similarity value a trigram language model has been used. TENOR contains 2 models both for English and Spanish texts, trained with the Brown corpus (Kucera and Francis, 1967) and the CESS-ESP (Martí and Taulé, 2007) respectively, with smoothing techniques (Chen and Goodman, 1996). This task has been implemented with the NLTK NgramModel class (Bird, 2006) for determining the replacement that minimises the perplexity, taking the latter as a measure of model quality.

## 4. Evaluation and Results

This section describes the evaluation process and the analysis of the results obtained with TENOR. First, the used corpora is introduced in Section 4.1. Subsequently, TENOR evaluation is explained in Section 4.2. Finally, the obtained results are discussed in Section 4.3.

### 4.1. Corpus

Two different corpora extracted from Twitter have been used in the evaluation process. Twitter<sup>10</sup> is an on-line microblogging service that enables its users to send and read textual messages of up to 140 characters. Due to this space constrain and its informal nature it can be considered a good source of short and noisy texts. Han's Twitter dataset<sup>11</sup> has been used for English texts and, following the same tagging scheme, a hand-annotated corpus of 1000 Tweets texts has been used for Spanish results<sup>12</sup>. In both cases, tagged words are annotated as out of vocabulary (OOV), inside the vocabulary (IV) or non-processable (NO). Also, for each OOV word its canonic version is provided.

### 4.2. Evaluation

We have evaluated TENOR performance in terms of precision and recall (Tang et al., 2005) taking into account OOV detection and normalisation separately (see Table 4). The obtained results were matched against the gold standard described in 4.1.

### 4.3. Results

TENOR results improve state-of-the-art approaches, with a 92% and a 82% F1 in OOV detection and OOV normalisation respectively (see Table 5).

<sup>8</sup><http://github.com/amsqr/Spanish-Metaphone>

<sup>9</sup>[http://en.wiktionary.org/wiki/Appendix:English\\_internet\\_slang](http://en.wiktionary.org/wiki/Appendix:English_internet_slang)

<sup>10</sup><http://www.twitter.com>

<sup>11</sup><http://www.csse.unimelb.edu.au/research/lt/resources/lexnorm/>

<sup>12</sup><http://gplsi.dlsi.ua.es/gplsi11/content/twitter-norm-dataset>

Raw Spanish		Normalised Spanish	
a)	tdo StO no s cierT, stams caNsa2	todo esto no es cierto, estamos cansados	
b)	xfa apoyo xa 1 niño d 3 añits	por favor apoyo para 1 niño de 3 años	
c)	mal momemto para sufrur!	mal momento para sufrir!	
d)	bamos a x ellos nesecitamos el apollo!!!!	vamos a por ellos necesitamos el apoyo!	
e)	amunt! valencia, visca el barça!	aumento! Valencia, busca el F.C. Barcelona!	
f)	el no aprobara	el no aprobara	
Raw English		Normalised English	
g)	whn ur talking to some1 an u say tht	When you are talking to someone and you say that.	
h)	Talkin abt this wee lasi cawd sophie:(	Talking about this week lazy caw sophie I'm sad.	
i)	WAAAAAAAY up great!	Way up great!	
j)	its my last wish to see u plz	Its my last wish to see you please.	

Table 3: Raw and normalised pairs of Spanish and English Web 2.0 examples.

	(OOV)	(IV)
Found	A	B
Not Found	C	D
<b>Precision:</b>	$P=A/(A+B)$	
<b>Recall:</b>	$R=A/(A+C)$	
<b>F1:</b>	$F=2PR/(P+R)$	

Table 4: Evaluation measures used in this study.

Task	Precision	Recall	F1
TENOR Eng. OOV	<b>91.7%</b>	<b>95.2%</b>	<b>93.4%</b>
TENOR Sp. OOV	<b>82.7%</b>	<b>98%</b>	<b>89.7%</b>
Han-Baldwin2011 OOV	61.1%	85.3%	71.2%
Han-Baldwin2011	75.3%	<b>75.3%</b>	75.3%
TENOR Eng.	88.9%	55.3%	68.2%
TENOR Eng. w/except.	<b>91.2%</b>	74.5%	<b>82.1%</b>
TENOR Sp.	94.1%	56%	70.2%
TENOR Sp. w/except.	<b>96.1%</b>	73%	<b>83%</b>

Table 5: Evaluation of out of vocabulary detection and normalisation results with and without the exception dictionary for English and Spanish Twitter texts.

Taking into account the obtained results, the use of the exception dictionary significantly enhances the normalisation of both English and Spanish texts. It can be noticed that Spanish normalisation results are higher, although its dictionary contains less entries than the English one. This is directly related to the results obtained in the OOV detection, in which the Spanish version of TENOR obtained slightly lower results (Fmeasure) than its English version. We can conclude that in Spanish is more difficult to detect OOV words and this is because there is a greater number of exceptions. By using the exception dictionary for English results also were improved, but this was expected since the exception dictionary was of greater size.

The obtained results show the capability of TENOR as a

tool for improving Web 2.0 texts accessibility by facilitating the work of current NLP simplification tools. Moreover, it also makes user-generated content more accessible to people unfamiliar with these text types.

## 5. Conclusions and Future Work

In this paper we have presented TENOR, a multilingual text normalisation approach for Web 2.0 texts. We have demonstrated that with TENOR noisy and difficult to understand English and Spanish texts can be converted into their canonic form. The substitution of non-normative vocabulary present in Web 2.0 texts, results in texts easier to understand and therefore makes the Web 2.0 new textual genres more accessible to everybody. This is a first step to facilitate the understanding of texts by easing the access to information using the new forms of communication available in the Web 2.0.

In the future we plan to extend TENOR by adding support to additional languages. Moreover, as a long term goal we would like to integrate this tool with text simplification strategies.

## Acknowledgements

This paper has been partially supported by Ministerio de Ciencia e Innovación - Spanish Government (grant no. TIN2009-13391-C04-01), Conselleria d'Educación - Generalitat Valenciana (grant no. PROMETEO/2009/119, ACOMP/2010/286 and ACOMP/2011/001) and the European Commission under the Seventh (FP7 - 2007-2013) Framework Programme for Research and Technological Development through the FIRST project (FP7-287607). This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

## 6. References

- Aiti Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for sms text normalization. *Proceedings of the COLING/ACL*, pages 33–40.
- John Barnden. 2008. Challenges in natural language processing: the case of metaphor (commentary). *International Journal of Speech Technology*, 11:121–123.

- Or Biran, Samuel Brody, and Noemie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Arnaldo Candido, Jr., Erick Maziero, Caroline Gasperin, Thiago A. S. Pardo, Lucia Specia, and Sandra M. Aluisio. 2009. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, pages 34–42, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-1996)*, pages 310–318.
- Walter Daelemans, Anja Höthker, and Erik Tjong Kim Sang. 2004. Automatic sentence simplification for subtitling in dutch and english. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048.
- Richard J. Evans. 2011. Comparing methods for the syntactic simplification of sentences in information extraction. *Literacy and Linguist Computing*, 26(4):371–388.
- Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011. Contextual Bearing on Linguistic Variation in Social Media. *ACL Workshop on Language in Social Media (LSM)*.
- 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, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Matt Huenerfauth, Lijun Feng, and Noemie Elhadad, 2009. *Comparing evaluation techniques for text readability software for adults with intellectual disabilities*, pages 3–10. ACM.
- Henry Kucera and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI, USA.
- Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. 2011. Insertion, deletion, or substitution? normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 71–76, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Veronica López, Rubén San-Segundo, Roberto Martín, Julian David Echeverry, and Syaheera Lutfi. 2010. Sistema de traducción de lenguaje SMS a castellano. In *XX Jornadas Telecom I+D*, Valladolid, Spain, September.
- María Antonia Martí and Mariona Taulé. 2007. Cess-ecce: corpus anotados del español y catalán. *Arena Romanistica. A new Nordic journal of Romance studies*, 1.
- Marcelo Mendoza, Barbara Poblete, and Carlos Castillo, 2010. *Twitter Under Crisis: Can we trust what we RT?*, volume 1060, page 10. ACM Press.
- Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners : A corpus analysis. *Electrical Engineering, (SLaTE)*:69–72.
- Lawrence Philips. 2000. The double metaphone search algorithm. *C/C++ Users Journal*, 18:38–43, June.
- John W. Ratcliff and David E. Metzner. 1988. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46–72, July.
- Horacio Saggion, Elena Gómez-Martínez, Alberto Anula, and Esteban Bourg, Lorena an dEtayo. 2011. Text simplification in simplex: Making texts more accessible. *Procesamiento del Lenguaje Natural*, 47:341–342.
- Daniel Scamfeld, Vanessa Scamfeld, and Elaine L Larson. 2010. Dissemination of health information through social networks: twitter and antibiotics. *American Journal of Infection Control*, 38(3):182–188.
- Claude. E. Shannon. 1948. A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:379–423.
- Jie Tang, Hang Li, Yunbo Cao, and Zhaohui Tang. 2005. Email data cleaning. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 489–498, New York, NY, USA. ACM Press.
- Sandra Williams and Ehud Reiter. 2008. Generating basic skills reports for low-skilled readers\*. *Nat. Lang. Eng.*, 14:495–525, October.