Spell-Checking in Spanish: The Case of Diacritic Accents

Jordi Atserias, Maria Fuentes, Rogelio Nazar, Irene Renau

Fundació Barcelona Media, TALP (Universitat Politècnica de Catalunya), IULA (Universitat Pompeu Fabra)
Av. Diagonal 177, 08018 Barcelona; C/Jordi Girona 1-3, 08034 Barcelona; Roc Boronat 138, 08018 Barcelona
jordi.atserias@barcelonamedia.org, mfuentes@lsi.upc.edu,{rogelio.nazar, irene.renau}@upf.edu

Abstract

This article presents the problem of diacritic restoration (or diacritization) in the context of spell-checking, with the focus on an orthographically rich language such as Spanish. We argue that despite the large volume of work published on the topic of diacritization, currently available spell-checking tools have still not found a proper solution to the problem in those cases where both forms of a word are listed in the checker's dictionary. This is the case, for instance, when a word form exists with and without diacritics, such as *continuo* 'continuous' and *continuó* 'he/she/it continued', or when different diacritics make other word distinctions, as in *continúo* 'I continue'. We propose a very simple solution based on a word bigram model derived from correctly typed Spanish texts and evaluate the ability of this model to restore diacritics in artificial as well as real errors. The case of diacritics is only meant to be an example of the possible applications for this idea, yet we believe that the same method could be applied to other kinds of orthographic or even grammatical errors. Moreover, given that no explicit linguistic knowledge is required, the proposed model can be used with other languages provided that a large normative corpus is available.

Keywords: computer-assisted writing in Spanish, diacritic restoration, n-gram language models, spell-checking

1. Introduction

Spell-checking is becoming increasingly important for the editorial industry as well as for end-users of word processors. Publishing companies, the press, scientists, teachers and a variety of other professionals and students (both native and second language speakers) are using them on a daily basis, thus making it one of the most widely used Natural Language Processing (NLP) technologies today. In addition to this, web and user-generated content (blogs, social media and the like) is growing at a brisk pace, making extensive use of non-standard language such as emoticons, spelling errors, letter casing, unusual punctuation and so on. As a consequence, spell-checking and text normalization – or restoration – are becoming even more important today, because this content cannot be processed correctly with general NLP tools unless all these different abnormalities are handled properly.

In this article we will focus on diacritic errors, that is to say, a single type of orthographic error which has not yet been satisfactorily solved by the most widely known spell-checking software applications. Diacritic marks are used in a large variety of languages and writing systems. In the case of Romance languages, diacritics are typically used to graphically mark the phonological accent, but can also be used for other purposes depending on the language and the lexical unit. In French, for instance, it is used in *é* or *è* to indicate how the vowel must be pronounced.

Errors involving accents can be trivially solved when there is only one correct possibility in a dictionary, such as in the case of the word *político* ('politician'). Confronted with an instance such as **politico*, it is easy for a spell-checking application to propose the correct word form, for instance by using an orthographic similarity measure. However, there can be other more interesting cases when the error gives rise to some kind of ambiguity (that is, two correct sentences with different meanings), because this can be particularly difficult to solve with automatic means.

In the case of Spanish, Catalan, French and other languages, diacritic accents are used in very common words to distinguish meaning and part-of-speech. Consider, for instance, the following Spanish sentences:

(1a.) No tenía *que* comer.
('S/he was not supposed to eat')

(1b.) No tenía *qué* comer.
('S/he had nothing to eat')

As shown in the translations, the accent in *qué* conveys an important distinction in meaning between sentences (1a) and (1b). This is the use traditionally linked to the term *diacritic* in Spanish (RAE, 2012, pp. 230-231)[1]. In this article, however, we use the term *diacritic* in a more general sense to refer to all the cases of words which have no other orthographic difference apart from the accent, as in the following cases:

(2a.) El agua del río está en *continuo* movimiento.
('The water of the river is in constant movement')

(2b.) *Continúo* con mi explicación.
('I will continue my explanation')

(2c.) *Continuó* con sus bromas todo el día.
('S/he continued with her/his jokes all day')

In (2.a), *continuo* ('continuous') is an adjective, whereas in (2.b) and (2.c) *continúo* and *continuó* are different tenses of the verb *continuar* ('to continue'). All these kinds of orthographic differentiations are the cause of frequent spelling mistakes even among well-educated speakers, who often forget to put the accent on a word or they do so incorrectly.

[1] The term refers to a closed list of pairs of words which are orthographically distinguished by the accent, which thus indicates their different grammatical category – e.g. *más* 'more' vs. *mas* 'but', *aún* 'still' vs. *aun* 'even', etc.

From this perspective, the problem that we address in this paper can be seen as a particular kind of ambiguity resolution. Our proposal to solve this problem is based on the idea that statistical models can be used as a simple and scalable technique to detect and solve a large proportion of the errors. Even considering a reduced context window, such as one word to the left and to the right of the target position, the frequency of the combinations of words in a large normative corpus can be taken as a clue to find the word the author really intended to produce. This makes a very simple and cost-effective solution as regards both implementation and computational effort. Moreover, the proposed models are language-independent and do not need any specific corpora other than well-written text (without lemmatization or POS-tagging). Despite the simplicity of the method, results show that it can be more effective than other more complex systems based on syntactic rules.

The remainder of the paper is organized as follows: the next section offers some brief comments on related work dealing with the subject of spell and grammar checking in general and the problem of diacritic restoration in particular. Section 3 explains the methodology proposed, while Section 4 presents the experimental set-up and the results. Finally, Section 5 draws some conclusions and discusses future work.

2. Related Work

The problem of spelling and grammar checking has been a research topic for decades in computational linguistics (Kukich, 1992). The first attempts were based on the idea that a combination of lexicon and hand-crafted grammar rules could suffice to solve the problem, both for spelling and grammar checkers (Heidorn et al., 1982; Cherry and McDonald, 1983), but later the interest shifted towards statistics-oriented methods (Angell et al., 1983; Atwell, 1987; Golding and Roth, 1999; Reynaert, 2004). Recently, different authors have begun to use the Web as a "normative" corpus (Moré et al., 2004; Hermet et al., 2008; Yi et al., 2008; Gamon et al., 2009). N-gram models of language have been used for grammar checking in German (Nazar, forthcoming) and Spanish (Nazar and Renau, forthcoming), as well as to identify collocation errors in L2 learners' compositions and other writings (Ferraro et al., 2011). In parallel, the problem of user-generated content (SMS, social media) has constituted a particular field of research on automatic text correction (Kobus et al., 2008).

The specific problem of diacritic restoration has been a topic of study in its own right. Seminal work in this field was conducted by Yarowsky (1994), who first outlined the problem of accent restoration as a disambiguation task. His approach combines an n-gram POS-tagger and a Bayesian classifier.

The POS-tagging approach (Hidden Markov Models) was also the first option in other early studies in French (El-Bèze et al., 1994; Simard, 1998; Simard and Deslauriers, 2001) and Romanian texts (Tufiş and Chitu, 1999; Tufiş and Ceauşu, 2007; Tufiş and Ceauşu, 2008). Mihalcea (2002) applied a grapheme-based approach, since character level n-grams apparently offer good performance as predictors of correct diacritics, not only in Romanian but also in a variety of other European languages. Among other advantages, such as being computationally efficient, grapheme methods are easy to apply to other languages. Building on this idea, promising results have been obtained in a variety of African languages, such as Cilubà, Gĩkũyũ, Kĩkamba and others (Wagacha et al., 2006; Pauw et al., 2007).

Research in automatic diacritic restoration has also been conducted in Urdu (Raza and Hussain, 2010), Sindhi (Mahar et al., 2011), Vietnamese (Nguyen and Ock, 2010), Māori (Cocks and Keegan, 2011) and Czech (Kanis and Müller, 2005), among other languages. In a different direction, which we suspect will generate more research, attempts have been made to restore accents in specialized terminology, because such terms often consist of out-of-vocabulary units, thereby ruling out any lexicon- or POS-tag-based approach (Zweigenbaum and Grabar, 2002).

The problem of diacritization is far more difficult in languages where diacritics play a fundamental role in distinguishing word senses (e.g. in Arabic, without diacritics the same word *k-t-b* could mean 'writer', 'book' or 'to write', among other possibilities). Some authors have applied the POS-tagging solution to diacritization in Arabic based on maximum entropy models (Zitouni et al., 2006; Mohamed and Kübler, 2009) and weighted finite-state transducers (Nelken and Shieber, 2005). Other authors have preferred to cast the problem of diacritization as a machine translation problem (Diab et al., 2007; Schlippe et al., 2008). Some have emphasized the importance of hybrid systems (Rashwan et al., 2009; Shaalan et al., 2009), while others, in contrast, prefer statistical models (Elshafei et al., 2006; Alghamdi et al., 2010).

As a general comment on the experiments that have been reported, we can see that there is still no consensus on how to evaluate the results of the experiments. Different evaluation methodologies on different datasets render incommensurable evaluation figures. Moreover, there are many factors that can artificially increase precision, such as counting the number of errors per total number of words in running text, because many words in the text do not contain any ambiguity and therefore would always count as correct. In other cases, such ambiguity can be negligible and thus picking the most frequent option already produces positive figures.

The main contribution of this paper to the subject of diacritic restoration is a new methodology based on word bigrams – which is computationally simple and language independent – and an empirical evaluation with a large volume of data. In the absence of a standardized method for evaluating this specific type of techniques, we also propose a method to prevent the biasing effect of high-frequency words, where the ambiguity is relatively easy to solve (because selecting the most frequent will probably be the right choice and this may lead to an overestimation of results). Finally, to our knowledge, since Yarowsky's (1994) study, no substantial work has been conducted on this subject in Spanish, apart from the some isolated attempt, such as (Galicia-Haro et al., 1999).

3. Methods

As already explained in the introduction, some of the spelling errors involving accents in Spanish can be solved by checking the words in a dictionary. However, very often this is not the case. When different accentuations are possible, they may also produce different meanings for the word (e.g. *médico* 'physician', *medicó*, 'he/she administered medication', *medico* 'I administer medication'). Henceforth, we name these sets of words "diacritic sets" and, in this paper, we will concentrate on these "diacritic" errors, i.e. those which cannot be trivially solved with a lexicon. Thus, the experiments we conducted first involved compiling corpora of these cases and then attempting to automatically correct them.

Lexical co-occurrence has been used in many NLP tasks, such as word choice in machine translation (Edmonds, 1997). In our case, we will use it to decide which of a series of differently accentuated variants is more likely in a given context. For our experiments, we will combine bigrams of the target word with the following and preceding words, backed-off with the unigrams of the target word[2].

Taking w_{target} as the target word form, w_{prev} as the word preceding w_{target}, w_{next} as the word following the target and w_1, w_2, w_n as the diacritic set associated to w_{target}, we define the following diacritic restoration methods:

- *baseline*: we will use a simple unigram model. This model will select the most frequent word form from among all the alternatives w_i, according to the corpus, as explained by Yarowsky (1994).
- *bigram*: will choose $\arg\max_{w_i} count(w_{prev}w_i) + count(w_{next}wi))$
- *bi+unigram*: will combine both strategies, i.e. choosing *bigram* if $\exists i \mid count(w_{prev}w_i) > 0$, otherwise choosing the most frequent word form as in the case of the baseline.

Following these strategies, three different models were built (baseline, bigram, bi+unigram) using 70% of a corpus of Spanish newspaper articles which comprises approximately 250 million tokens. The remaining 30% is used as a test set for experiments in Section 4.2.

4. Results

Evaluating performance on a diacritic restoration task is complex for a number of different reasons. Obtaining a corrected corpus of real human errors is costly and, in addition, such a sample may bias the evaluation to a particular setting, because error types may vary greatly depending on the type of people involved (e.g. native speakers, second language learners or media-related errors, such as mobile mistypings and so on). In order to account for this variation, we performed two kinds of evaluations, first on a small corpus of real human errors (Section 4.1.) and then on a large corpus of automatically generated diacritic errors (Section 4.2.).

Method	Precision	Recall	F1
bi+unigram	0.82	**0.82**	**0.82**
bigram	**0.83**	0.79	0.81
baseline	0.65	0.65	0.65
Microsoft Word 2007	0.31	0.31	0.31
Stilus	0.30	0.30	0.30
Google Docs	0.01	0.01	0.01
correctorortografico.com	0.01	0.01	0.01

Table 1: Full results for the 100 sentences containing common errors from student materials

4.1. Results on a reduced-scale sample of real errors

In order to evaluate the accuracy of the proposed method in detecting and solving real errors, we collected 100 sentences from text materials written by native Spanish-speaking students, each of which contained one diacritic error. The source for this material was 'El Rincón del Vago' (http://www.rincondelvago.com), a popular website among primary and secondary school pupils that is used to share notes, essays and other academic writings. As a consequence of the fact that the majority of these texts have not been submitted to any kind of revision by the teacher, they frequently contain various grammar and spelling errors. Some examples of sentences with diacritic problems follow:

(3a.) Aunque no era soldado **participo* en varias guerras.
('Despite the fact that he was not a soldier, he nevertheless participated in different wars.')

(3b.) En **éste* fragmento de texto...
('In this fragment of text...')

(3c.) Solo pensaba en **si* mismo.
('[He] only used to care about himself.')

In (3a), instead of *participó* ('[he] participated'), the student chose the wrong form **participo*, which is the first person simple present of the same verb. In (3b), the accent was added incorrectly in **éste fragmento* (instead of *este fragmento*), an ungrammatical combination in which the determiner *este* ('this') is confused with the pronoun *éste* ('this one'). Finally, in 3c the reflexive pronoun *sí* ('him/herself') is confused with the conjunction *si* ('if').

Table 1 shows the result of applying the different models described in the previous section to this sample of real accent errors, along with the results obtained with different commercial spelling and grammar checkers, including both document processors and web services, applied to the same sample. Since commercial spell-checkers always take a decision (correct/incorrect), precision and recall are the same and thus F1. Regarding the proposed methods, bigrams alone cannot always take a decision since not all bigrams are necessarily present in the training corpus, thus producing different figures for precision and recall. At the Unigram level, in contrast, there is a much higher probability of occurrence of some of the word forms in the diacritics set.

[2] In future work we may use scalable implementations of n-gram language models (Pauls and Klein, 2011)

Therefore, it is more likely that a decision will be taken. It should also be noticed that, in this experimental set-up, the word forms in the diacritic sets are not necessarily a closed list (i.e. there could be new forms in the test set).

As explained in Section 3, the baseline consists in selecting the most frequent word, which can be deemed to be the simplest solution to the problem and yet it largely outperforms the commercial spell-checkers. The difference in the results of our algorithm with respect to the other commercial spell-checkers is also highly significant (0.5 increase in performance).

A qualitative analysis of the errors enables us to gain some insight into the limitations and potential of the different systems. MS Word gives correct solutions for (3c), but not for (3a) and (3b) (the opposite to the proposed models). A probable reason for this may be that, for a rule-based system, it is easier to create a rule for *sí mismo* ('himself') than others for verbs, such as in (3a), although it is surprising that it was not able to give a solution for **éste* fragmento, the combination of the pronoun *éste* plus a noun being clearly ungrammatical. Stilus offers the right correction in (3b) and (3c), but fails in (3a).

With respect to our method, the bi+unigram model is able to detect errors such as *Fred le *pregunto* (instead of *Fred le preguntó*, 'Fred asked him/her'), in which *preguntó* is the third person singular of the simple past and *pregunto* is the first person singular of the simple present of the verb *preguntar* ('to ask'). The system can make this correction because the third person is far more frequent in the corpus in a context such as *le pregunto/le preguntó*. By the same token, the model is right when correcting *recibieron *ordenes* (instead of *recibieron órdenes*, 'they received orders'), in which *ordenes* ('that you order') is confused with *órdenes* ('orders'): in this case, the combination of two verbs like *recibieron* and *ordenes* is ungrammatical and it does not occur in the corpus. Nevertheless, the algorithm does not detect mistakes such as *cuando *esta nuevamente en el frente* ('when [he] is in the front again'), in which the verb *está* ('[he/she] is') is confused with the determiner *esta* ('this', feminine). This is because the context *cuando esta/está nuevamente* is equally possible for both cases. In the three sentences given as examples in (3), bi+unigrams proposed the right correction in (3a) and (3b) but failed in (3c).

In our opinion, the variety of results offered by the different systems reinforces the idea that a combination of statistics and rule-based checkers could be an optimal solution for the problem at hand.

4.2. Results on large-scale artificial errors

To avoid the cost of manually collecting a large corpus of real orthographic errors, we built a corpus of artificial accent errors derived from a lexicon by simply comparing the "unaccented" version of correct words. Following this method, 178,596 words ordered in 89,109 diacritic sets were extracted from a Spanish lexicon. With these sets of words with different accents, it is possible to build training and evaluation datasets. Every time one of these words is found, we obtain a positive example, and by replacing the word by any other member of its diacritic set, we obtain artificial negative examples. For instance, given the correct sentence (4a), and considering the diacritic set {*esta-está-ésta*}, three example sentences with different labels can be generated: the correct original sentence (4a) plus two incorrect examples, as shown in (4b) and (4c).

(4a.) Ella *está* en casa cuando Javier llega del colegio.
('She is at home when Javier arrives from school.')

(4b.) Ella **esta* en casa cuando...

(4c.) Ella **ésta* en casa cuando...

Method	Precision	Recall	F1
baseline	0.96	0.96	0.96
bi+unigram	**0.98**	0.94	0.96
bigram	**0.98**	**0.98**	**0.98**

Table 2: Results on the artificial error corpus

On applying this methodology to the 30% of the newspaper corpus that we did not use for training, we obtained an artificial error corpus containing 12,805,205 examples. Results are shown in Table 2, where it can be seen that the bigram model is able to solve most of the problems correctly, but also that the most frequent word (i.e. the baseline method) performs almost equally well. Two factors might explain these results:

1. Some of the sets are more frequent than others.
2. The frequency in the artificial error corpus does not necessarily correspond to the frequency of human writing errors.

Consequently, and in order to further explore the impact of word frequency in our experiments, we removed the "easiest" cases from the diacritic sets (i.e. those where the relative frequency of one word is higher than 70%, and thus the most frequent would have a precision higher than 0.7). The resulting test corpus is smaller (1.146.015 words) and the results of replicating the experiment on such a dataset, shown in Table 3, indicate a lower performance pattern. This is consistent, however, with those figures reported in the experiment with real errors (Table 4.1.), which suggests that this is the most reliable estimation of performance for our method.

Method	Precision	Recall	F1
baseline	0.61	0.60	0.61
bi+unigram	0.84	**0.84**	**0.84**
bigram	**0.85**	0.64	0.73

Table 3: Results on the filtered corpus

5. Conclusions and Future Work

Diacritics represent a spelling problem of considerable importance in Spanish as well as in other languages. This paper addresses a spell-checking problem in Spanish which has still not been fully solved by today's text processors.

The difficulty of automatic accent correction in Spanish resides in the fact that changes in the accentuation of a word often produce different words that do exist but have different meanings, and it is on cases like these that our work is focused.

The experiment reported here shows that a simple n-gram technique can solve most of the diacritic errors. In addition, we have also provided a framework for a deep exploration of Spanish accentuation, since the list of diacritic sets and the artificial error corpus can also be a valuable aid in second language acquisition to better understand and access the Spanish accentuation rules.

As future work, we plan to conduct more extensive evaluations on Spanish and other languages. Attempts will also be made to extend the current framework to deal with other types of orthographic and typographic errors by including experiments with more than one error per sentence. Furthermore, we will try to expand the generalization power of our training corpus by replacing actual words by their corresponding semantic classes (using synonyms and hypernyms), thus converting a text corpus into a pattern corpus, as in Gross's (1994) theory of classes of objects (e.g. bigrams such as *vehículo aulló* 'vehicle howled' or *animal aulló* 'animal howled' are more general than *ambulancia aulló* 'ambulance howled' or *lobo aulló* 'wolf howled'). With this transformation, we hope that the corpus will be able to represent bigrams that are not actually instantiated within it.

6. Acknowledgements

This research has been made possible thanks to funding from projects 'Agrupación semántica y relaciones lexicológicas en el diccionario', lead researcher J. DeCesaris (HUM2009-07588/FILO); APLE: 'Procesos de actualización del léxico del español a partir de la prensa', lead researcher: M.T. Cabré (FFI2009-12188-C05-01/FILO); Holopedia (TIN2010-21128-C02-02) and KNOW-2 (TIN2009-14715-C04-04). The authors would like to thank the anonymous reviewers for their constructive remarks.

7. References

M. Alghamdi, Z. Muzaffar, and H. Alhakami. 2010. Automatic restoration of arabic diacritics: a simple, purely statistical approach. *The Arabian Journal for Science and Engineering*, 35(2c):125–135.

R. Angell, G. Freund, and P. Willett. 1983. Automatic spelling correction using a trigram similarity measure. *Information Processing & Management*, 19(4):255–261.

E. Atwell. 1987. How to detect grammatical errors in a text without parsing it. In *Proceedings of the Third Conference of the European ACL*, pages 38–45, Copenhagen.

L. Cherry and N. McDonald. 1983. The writers workbench software. *Byte*, pages 241–248.

J. Cocks and T. Keegan. 2011. A word-based approach for diacritic restoration in māori. In *Proceedings of Australasian Language Technology Association Workshop*, pages 126–130.

M. Diab, M. Ghoneim, and N. Habash. 2007. Arabic diacritization in the context of statistical machine translation. In *Proceedings of the Machine Translation Summit (MT-Summit)*, Copenhagen.

P. Edmonds. 1997. Choosing the word most typical in context using a lexical co-occurrence network. In *Proceedings of the eighth conference on European chapter of the ACL*, EACL '97, pages 507–509, Stroudsburg, PA, USA. ACL.

M. El-Bèze, B. Mérialdo, B. Rozeron, and A. Derouault. 1994. Accentuation automatique de texte par des méthodes probabilistes. *Technique et sciences informatiques*, 13(6):797–815.

M. Elshafei, H. Al-Muhtaseb, and M. Alghamdi. 2006. Statistical methods for automatic diacritization of arabic text. In *Proceedings of the Saudi 18th National Computer Conference (NCC18)*, Riyad.

G. Ferraro, R. Nazar, and L. Wanner. 2011. Collocations: A challenge in computer-assisted language learning. In *Proceedings of the 5th International Conference on Meaning-Text Theory*, pages 69–79, Barcelona.

S. Galicia-Haro, I. Bolshakov, and A. Gelbukh. 1999. A simple spanish part of speech tagger for detection and correction of accentuation error. In *Proceedings of the Second International Workshop on Text, Speech and Dialogue*, TSD '99, pages 219–222, London, UK. Springer-Verlag.

M. Gamon, C. Leacock, C. Brockett, W. Dolan, J. Gao, D. Belenko, and A. Klementiev. 2009. Using statistical techniques and web search to correct esl errors. *CALICO Journal*, 26(3):491–511.

A. Golding and D. Roth. 1999. A winnow-based approach to context-sensitive spelling correction. *Machine Learning - Special issue on natural language learning archive*, 34(1-3).

G. Gross. 1994. Classes d'objets et description des verbes. *Langages*, 115.

G. Heidorn, K. Jensen, L. Miller, R. Byrd, and M. Chodorow. 1982. The epistle text-critiquing system. *IBM Systems Journal*, (21):305–326.

M. Hermet, A. Dsilets, and S. Szpakowicz. 2008. Using the web as a linguistic resource to automatically correct lexico-syntactic errors. In *Proceedings of the Language Resources and Evaluation Conference (LREC 08)*, pages 874–878, Marrakesh.

J. Kanis and L. Müller. 2005. Using lemmatization technique for automatic diacritics restoration. In *Proceedings of SPECOM 2005*, pages 255–258, Moscow. Moscow State Linguistic University.

C. Kobus, F. Yvon, and G. Damnati. 2008. Normalizing sms: are two metaphors better than one? In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1 (COLING '08)*, pages 441–448. ACL.

K. Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys*, (24):377–439.

J. Mahar, G. Memon, and H. Shaikh. 2011. Sindhi diacritics restoration by letter level learning approach.

Sindh University Research Journal (Science Series), 43(2):119–126.

R. Mihalcea. 2002. Diacritics restoration: Learning from letters versus learning from words. In A. Gelbukh, editor, *CICLing 2002. LNCS*, volume 2276, pages 339–348.

E. Mohamed and S. Kübler. 2009. Diacritization for real-world arabic texts. In *Proceedings of the International Conference RANLP-2009*, pages 251–257, Borovets, Bulgaria, September. ACL.

J. Moré, S. Climent, and A. Oliver. 2004. A grammar and style checker based on internet searches. In *Proceedings of the Language Resources and Evaluation Conference (LREC 04)*, pages 1931–1934, Marrakesh.

R. Nazar and I. Renau. forthcoming. Google books n-gram corpus used as a grammar checker. In *Proceedings of the Second Workshop on Computational Linguistics and Writing (CL&W 2012), EACL 2012*, Avignon.

R. Nazar. forthcoming. Algorithm qualifies for c1 courses in german exam without previous knowledge of the language. In *Proceedings of the 6th Corpus Linguistics Conference, Birmingham 2011*, Birmingham.

R. Nelken and S. Shieber. 2005. Arabic diacritization using weighted finite-state transducers. In *Proceedings of the 2005 ACL Workshop on Computational Approaches to Semitic Languages*, pages 79–86, Ann Arbor.

Kiem-Hieu Nguyen and Cheol-Young Ock. 2010. Diacritics restoration in vietnamese: letter based vs. syllable based model. In *Proceedings of the 11th Pacific Rim international conference on Trends in artificial intelligence*, PRICAI'10, pages 631–636, Berlin, Heidelberg. Springer-Verlag.

Adam Pauls and Dan Klein. 2011. Faster and smaller n-gram language models. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *ACL*, pages 258–267. The Association for Computer Linguistics.

G. De Pauw, P. Wagacha, and G. De Schryver. 2007. Automatic diacritic restoration for resource-scarce languages. In V Matousek and P Mautner, editors, *Lecture Notes in Artificial Intelligence*, volume 4629, pages 170–179. Springer.

RAE. 2012. *Ortografía de la lengua española*. (Real Academia Española). Espasa Calpe, Madrid, 1st edition.

M. Rashwan, M. Al-Badrashiny, M. Attia, and S. Abdou. 2009. A hybrid system for automatic arabic diacritization. In Khalid Choukri and Bente Maegaard, editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt. The MEDAR Consortium.

A. Raza and S. Hussain. 2010. Automatic diacritization for urdu. In *Proceedings of the Conference on Language and Technology 2010 (CLT10)*, pages 105–111, Islamabad.

M. Reynaert. 2004. Text induced spelling correction. In *Proceedings of COLING 2004*, pages 1931–1934, Geneva.

T. Schlippe, T. Nguyen, and S. Vogel. 2008. Diacritization as a machine translation problem and as a sequence labeling problem. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, pages 270–278, Waikiki.

K. Shaalan, H. Abo Bakr, and I. Ziedan. 2009. A hybrid approach for building arabic diacritizer. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Semitic '09, pages 27–35, Stroudsburg, PA, USA. ACL.

M. Simard and A. Deslauriers. 2001. Realtime automatic insertion of accents in french text. *Natural Language Engineering*, 7(2):143–165.

M. Simard. 1998. Automatic insertion of accents in french text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP-2.

D. Tufiş and A. Ceauşu. 2007. Diacritics restoration in romanian texts. In *RANLP 2007 Workshop: A Common Natural Language Processing Paradigm for Balkan Languages*, Borovets.

D. Tufiş and A. Ceauşu. 2008. Diac+: A professional diacritics recovering system. In *Proceedings of LREC 2008 (Language Resources and Evaluation Conference)*, Marakkech. ELRA.

D. Tufiş and A. Chitu. 1999. Automatic insertion of diacritics in romanian texts. In *Proceedings of the 5th International Workshop on Computational Lexicography COMPLEX*, pages 185–194, Pecs.

P. Wagacha, G. De Pauw, and P. Githinji. 2006. A grapheme-based approach for accent restoration in gikuyu. In *Proceedings of LREC 2006 (Language Resources and Evaluation Conference)*, pages 1937 – 1940, Marakkech. ELRA.

D. Yarowsky. 1994. A comparison of corpus-based techniques for restoring accents in spanish and french text. In *Proceedings of the 2nd Annual Workshop on very large Text Corpora*, pages 99–120, Las Cruces.

X. Yi, G. Gao, and W. Dolan. 2008. A web-based english proofing system for english as a second language users. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 619–624, Hyderabad.

I. Zitouni, J. Sorensen, and R. Sarikaya. 2006. Maximum entropy based restoration of arabic diacritics. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, ACL-44, pages 577–584, Stroudsburg, PA, USA. ACL.

P. Zweigenbaum and N. Grabar. 2002. Restoring accents in unknown biomedical words: application to the french mesh thesaurus. *International Journal of Medical Informatics*, 67:113–126.