5th International Conference on Corpus Linguistics (CILC2013)

# esTenTen, a Vast Web Corpus of Peninsular and American Spanish

Adam Kilgarriff[a]*, Irene Renau[b]

*aLexical Computing Ltd., 71 Freshfield Road, Brighton BN2 0BL, UK*
*bInstitut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, C/Roc Boronat 138, 08018 Barcelona, Spain*

## Abstract

Everyone working on general language would like their corpus to be bigger, wider-coverage, cleaner, duplicate-free, and with richer metadata. As a response to that wish, Lexical Computing Ltd. has a programme to develop very large 'TenTen' web corpora. In this paper we introduce the Spanish corpus, esTenTen, of 8 billion words and 19 different national varieties of Spanish. We investigate the distance between the national varieties as represented in the corpus, and examine in detail the keywords of Peninsular Spanish *vs.* American Spanish, finding a wide range of linguistic, cultural and political contrasts.

*Keywords:* corpus linguistics, Sketch Engine, Spanish dialects, TenTen corpora

## 1. The TenTen family of corpora

Everyone working on general language would like their corpus to be bigger, wider-coverage, cleaner, duplicate-free, and with richer metadata. As a response to that wish, Lexical Computing Ltd. has a programme to develop the 'TenTen' family of corpora.[1] It builds on the projects described in Sharoff (2006) and Baroni et al. (2009) to build very large web corpora. At time of writing, multi-billion-word corpora have been built for Arabic, Chinese, English,

_____

* Corresponding author. Tel.: +44 7971867845
  *E-mail address:* adam@lexmasterclass.com

[1] Regarding the name: the WaC ("Web as Corpus") suffix has been widely used by web corpus builders, and many corpora have names comprising the language-specific code followed by "WaC". To forestall confusion with a name like FrWaC being ambiguous between two different corpora (both French and web-crawled) a new name was needed. The new batch of corpora are in the order $10^{10}$ (10 billion) words, so this is the TenTen family.

French, German, Italian, Japanese, Korean, Portuguese, Russian and Spanish; they have been lemmatised, part-of-speech tagged and made available in the Sketch Engine. In this paper we introduce the Spanish corpus, esTenTen, and explore its composition in terms of national varieties.

## 2. The processing chain

The processing chain for creating a TenTen corpus is:

- Crawl the web with Spiderling (Pomikalek & Suchomel 2012), a crawler designed specifically for preparing linguistic corpora. For Spanish the crawl was restricted to those websites with a top level domain (the last part of the URL) of one of the Spanish-speaking countries.
- Remove non-textual material and boilerplate with jusText (Pomikalek 2011). JusText uses the working definition that we want only 'text in sentences' (and not, e.g. headers and footers). The algorithm is linguistically informed, rejecting material that does not have a high proportion of tokens that are the grammar words of the language, so, in the course of data-cleaning, most material which is not in the desired language is removed.
- De-duplicate with Onion (Pomikalek 2011). We de-duplicate at the paragraph level, as, for many linguistic purposes, a sentence is too small a unit, but a whole web page (which may contain large chunks of quoted material) is too large.

These tools are designed for speed and we use them installed in a cluster of servers. For a language like Spanish where there is plenty of material available, we can gather, clean and de-duplicate a billion words a day.

Then, we want to tokenize the corpus into words, lemmatise and part-of-speech tag. For these processes we examine the available tools for the language and apply the best we can find (after considering, firstly, accuracy, but also speed, quality of engineering, and licence terms). For Spanish we have explored both Freeling (Padró & Stalinovsky 2012) and TreeTagger (Schmid 1994): our comparison of the pros and cons of the two systems is ongoing. For the current exercise we have used a version of the corpus processed by TreeTagger.

## 3. esTenTen and national varieties

esTenTen is a corpus of 8.38 billion words, broken down between different national varieties (as identified by URL) as follows:

Table 1. National varieties in esTenTen.

| Country | Suffix | Words (millions) | Documents (thousands) |
| --- | --- | --- | --- |
| Argentina | .ar | 2,447 | 6,002 |
| Bolivia | .bo | 47 | 137 |
| Chile | .cl | 859 | 2,254 |
| Columbia | .co | 371 | 829 |
| Costa Rica | .cr | 47 | 114 |
| Cuba | .cu | 211 | 378 |
| Dominican Rep | .do | 43 | 132 |
| Ecuador | .ec | 64 | 231 |
| El Salvador | .sv | 27 | 69 |
| Guatemala | .gt | 27 | 80 |
| Honduras | .hn | 8 | 25 |
| Mexico | .mx | 1,470 | 3,543 |
| Nicaragua | .ni | 53 | 101 |
| Panama | .pa | 15 | 65 |

| Country | Suffix | Words (millions) | Documents (thousands) |
|---------|--------|------------------|------------------------|
| Paraguay | .py | 51 | 146 |
| Peru | .pe | 253 | 609 |
| Spain | .es | 1,992 | 4,377 |
| Uruguay | .uy | 156 | 358 |
| Venezuela | .ve | 218 | 602 |

As Baroni, Sharoff and colleagues have shown (Sharoff 2006, Baroni et al. 2009), web corpora give a useful and interesting picture of 'general language' and, where we can compare the picture of the language given by the web corpus with that given by some other reference corpus, the picture from the web corpus is, for many purposes, as good as or better than that given by the reference corpus (see also Kilgarriff 2012).[2]

One research topic where esTenTen shows potential is in exploring regional variation. We have large samples from nineteen different national top-level domains: it seems likely that these correspond to nineteen different regional variants of the language. While many Spanish-language documents from Uruguay will not be in the .ur top level domain, we think it unlikely there are many .ur documents that are not from Uruguay, so the .ur subcorpus will be of Uruguayan Spanish, and we can reasonably call it a national subcorpus. The Sketch Engine supports the creation and exploration of subcorpora, so, while searches and word lists from esTenTen contain, by default, Peninsular and Latin American Spanish, they can easily be restricted to one variety, or a subset of varieties.

As all the national subcorpora have been created using exactly the same method, it is plausible that they all contain the same mix of different types of texts – blog, newspaper, academic journal, sports report, club page, company report, personal home pages, etc. If this is so, then the keywords of any one national subcorpus will be the distinctive vocabulary of that national variety of Spanish, along with those words that are used elsewhere but are used particularly in that country. To the extent that the different national subcorpora contain different mixes of text types, the linguistic differences between the national varieties will be mixed in with differences following from the different proportions of text types. If, for example, the Uruguayan subcorpus has a particularly high proportion of corporate pages, we will probably find words like *empresa* ('company'), *corporación* ('corporate'), *beneficios* ('profit'), *dividendos* ('dividend') in the keyword list alongside distinctively Uruguayan words.

### 3.1. Quantitative comparison

One question of interest is: which varieties are most similar? We can explore this by computing distances between all the national subcorpora (Figure 1).

---

[2] While, for Spanish, there are two very large and carefully designed corpora which can be searched on the web, the "Corpus de Referencia del Español Actual" (CREA), from the Spanish Royal Academy, and the "Corpus del Español" from Mark Davies at Brigham Young University, in neither case did we have access to the full text. This meant we could not load them into the Sketch Engine and review corpus distance measures and keywords of each corpus *vs.* the other, for thoroughgoing comparison. We hope this may be possible in the future.

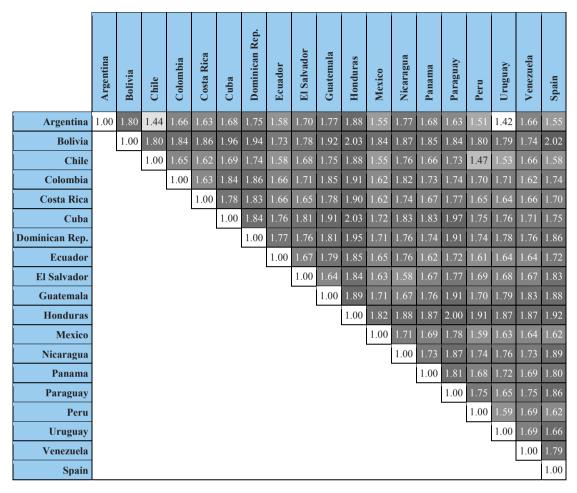| | Argentina | Bolivia | Chile | Colombia | Costa Rica | Cuba | Dominican Rep. | Ecuador | El Salvador | Guatemala | Honduras | Mexico | Nicaragua | Panama | Paraguay | Peru | Uruguay | Venezuela | Spain |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Argentina | 1.00 | 1.80 | 1.44 | 1.66 | 1.63 | 1.68 | 1.75 | 1.58 | 1.70 | 1.77 | 1.88 | 1.55 | 1.77 | 1.68 | 1.63 | 1.51 | 1.42 | 1.66 | 1.55 |
| Bolivia | | 1.00 | 1.80 | 1.84 | 1.86 | 1.96 | 1.94 | 1.73 | 1.78 | 1.92 | 2.03 | 1.84 | 1.87 | 1.85 | 1.84 | 1.80 | 1.79 | 1.74 | 2.02 |
| Chile | | | 1.00 | 1.65 | 1.62 | 1.69 | 1.74 | 1.58 | 1.68 | 1.75 | 1.88 | 1.55 | 1.76 | 1.66 | 1.73 | 1.47 | 1.53 | 1.66 | 1.58 |
| Colombia | | | | 1.00 | 1.63 | 1.84 | 1.86 | 1.66 | 1.71 | 1.85 | 1.91 | 1.62 | 1.82 | 1.73 | 1.74 | 1.70 | 1.71 | 1.62 | 1.74 |
| Costa Rica | | | | | 1.00 | 1.78 | 1.83 | 1.66 | 1.65 | 1.78 | 1.90 | 1.62 | 1.74 | 1.67 | 1.77 | 1.65 | 1.64 | 1.66 | 1.70 |
| Cuba | | | | | | 1.00 | 1.84 | 1.76 | 1.81 | 1.91 | 2.03 | 1.72 | 1.83 | 1.83 | 1.97 | 1.75 | 1.76 | 1.71 | 1.75 |
| Dominican Rep. | | | | | | | 1.00 | 1.77 | 1.76 | 1.81 | 1.95 | 1.71 | 1.76 | 1.74 | 1.91 | 1.74 | 1.78 | 1.76 | 1.86 |
| Ecuador | | | | | | | | 1.00 | 1.67 | 1.79 | 1.85 | 1.65 | 1.76 | 1.62 | 1.72 | 1.61 | 1.64 | 1.64 | 1.72 |
| El Salvador | | | | | | | | | 1.00 | 1.64 | 1.84 | 1.63 | 1.58 | 1.67 | 1.77 | 1.69 | 1.68 | 1.67 | 1.83 |
| Guatemala | | | | | | | | | | 1.00 | 1.89 | 1.71 | 1.67 | 1.76 | 1.91 | 1.70 | 1.79 | 1.83 | 1.88 |
| Honduras | | | | | | | | | | | 1.00 | 1.82 | 1.88 | 1.87 | 2.00 | 1.91 | 1.87 | 1.87 | 1.92 |
| Mexico | | | | | | | | | | | | 1.00 | 1.71 | 1.69 | 1.78 | 1.59 | 1.63 | 1.64 | 1.62 |
| Nicaragua | | | | | | | | | | | | | 1.00 | 1.73 | 1.87 | 1.74 | 1.76 | 1.73 | 1.89 |
| Panama | | | | | | | | | | | | | | 1.00 | 1.81 | 1.68 | 1.72 | 1.69 | 1.80 |
| Paraguay | | | | | | | | | | | | | | | 1.00 | 1.75 | 1.65 | 1.75 | 1.86 |
| Peru | | | | | | | | | | | | | | | | 1.00 | 1.59 | 1.69 | 1.62 |
| Uruguay | | | | | | | | | | | | | | | | | 1.00 | 1.69 | 1.66 |
| Venezuela | | | | | | | | | | | | | | | | | | 1.00 | 1.79 |
| Spain | | | | | | | | | | | | | | | | | | | 1.00 |

Fig. 1. Distances between national subcorpora. Lighter cells show smaller distances.

Our method takes the 500 commonest words across the two subcorpora, and then, for each word

- normalizes both frequency to 'per million words'
- adds a parameter of 100 to both of these numbers,[3] and
- divides the larger by the smaller.

We then sum the 500 scores derived in this way and divide by 500 to give a measure of the distance between the two corpora, which is an 'average ratio' for the difference between the two corpora, with a score of 1.00 if a corpus is compared with itself. Methods such as this are motivated, explored and evaluated in detail in Kilgarriff (2001), and this method is implemented in the Sketch Engine.

There are several observations to make on the figure. The lightest cells (excluding the leading diagonal) are, in order, for Argentina-Uruguay, Argentina-Chile, Chile-Peru, Argentina-Peru, Chile-Uruguay. These southernmost countries form a cluster with similar varieties (the three most similar pairs all have long borders), even though they

---

[3] This has the effect of giving more weight to higher-frequency words. For a full discussion see Kilgarriff (2009).

do not belong to the same variety of Spanish (except in the case of Argentina-Uruguay, both included in the Spanish variety of the River Plate region) –see Moreno Fernández and Otero Roth 2007. This is a matter worthy of further investigation.

The Peninsular variety shows differences with respect to other dialects, but also remarkable similarities which, with the data to hand, do not distinguish it from the American varieties.

The darkest cells are for Honduras, Bolivia and Paraguay. All of these countries had comparatively small subcorpora, and the greater distances may well be an artifact of the distance measure (there is a tendency to get higher scores for smaller corpora, which we are currently investigating) though in Bolivia and Paraguay it may also be related to the strong presence of indigenous languages. Conversely, the most consistently pale cells are for Mexico, which has the third largest volume of data.

### 3.2. Qualitative comparison: European vs. Latin American Spanish

Using the keyword function of the Sketch Engine, we identified the 100 'most Spanish' and 'most American' words:

**American, first 100 keywords:**
*nos, nuestra, quienes, nuestros, federal, mis, argentino, mexicano, diputado, peso, costo, agregar, nuestras, gobernador, gobierno, estado, nacional, sostener, brindar, autoridad, cual, lograr, escuela, intendente, chileno, región, dirigente, auto, funcionario, presidente, país, señalar, entregar, miles, capacitación, policía, comuna, plata, senador, luego, institución, distrito, dólar, justicia, república, estatal, legislador, ley, cancha, expresar, aporte, cruz, mil, promedio, generar, candidato, provincia, ingresar, vos, manejo, político, ubicar, integrante, monto, municipio, acá, mundial, municipalidad, municipal, manifestar, productor, indígena, estudiante, constitucional, evento, cargo, mencionar, torneo, salud, docente, reclamo, fortalecer, penal, cubano, involucrar, nuevamente, armar, electoral, secretario, plantel, mandatario, juez, dios, tomar, terminar, venezolano, tribunal, área, enfrentar, oportunidad*

**Peninsular, first 100 keywords:**
*esto, era, euro, nosotros, ayuntamiento, español, apartado, situar, coche, coste, hotel, nuestro, suponer, recoger, consejería, actuación, conseguir, acabar, web, europeo, ordenador, habitación, vuestro, vídeo, colaboración, celebrar, subvención, plaza, vosotros, importe, disponer, añadir, blog, financiación, gustar, autonómico, formación, coger, entorno, resolución, haber, página, asignatura, curso, catalán, consejero, tienda, mejora, encantar, edición, usuario, facilitar, dato, habitual, proceder, peseta, relativo, restaurante, película, bastante, prever, socialista, aunque, ámbito, empleo, autónomo, alquiler, madrileño, conceder, acoger, competición, premio, disfrutar, ofertar, andaluz, ofrecer, echar, anexo, aplicación, viajero, número, convocatoria, fase, plazo, tras, también, enlace, aportación, concurso, valenciano, verano, profesional, antiguo, siguiente, online, tráfico, adaptar, implantación, sanitario, título*

We specified lemmas rather than word forms, and restricted results to all-lower-case items of three or more letters, since names (which dominate capitalised items) were not of great interest, and in our experience one and two letter items include many abbreviations and acronyms which are hard to interpret. We used the parameter of 100 (see previous footnote). As already noted, we expect these lists to contain some words that are there because of dialectal differences, and others that are there because of different proportions of different text types. While the first set of items are of direct linguistic interest, the second set are worthy of careful consideration as well, as they may tell us about cultural differences between the Spanish-speaking parts of Europe and America, and may also serve to warn us about problematic differences between the two subcorpora. (For a full discussion, see Kilgarriff 2012.)

We have looked closely at these lists, putting words in similar linguistic classes or semantic fields together, and assembled a classification of the differences between the varieties as represented by their subcorpora. For the American subcorpus, some are words commonly used in more than one American country, and others that are in this list because they are very common in just one or two countries. Table 2 presents our classification.

Table 2. A classification of American and Peninsular Spanish keywords in esTenTen.

| | **American esTenTen** | **Peninsular esTenTen** |
|---|---|---|
| Lexical preferences | aporte, evento, lograr, manejo, reclamo, terminar, ubicar | acabar, acoger, alquiler, conseguir, habitación, recoger, situar, tienda |
| Geographical adjectives | argentino, chileno, cubano, mexicano, venezolano | andaluz, catalán, español, europeo, madrileño, valenciano, |
| Administrative divisions | **comuna**, distrito, estado, federal, **municipalidad**, municipio, provincia | **ayuntamiento** |
| Politics and Administration | autoridad, diputado, candidato, cargo, dirigente, electoral, funcionario, gobernador, gobierno, estatal, institución, **intendente**, mandatario, municipal, nacional, país, policía, político, presidente, región, república, secretario, senador | anexo, apartado, autonómico, autónomo, consejería, consejero, convocatoria, implantación, plazo, resolución, socialista, subvención |
| Grammatical words | acá, cual, luego, mis, nos, nuestra, nuestras, nuestros, quienes, **vos** | aunque, bastante, echar, esto, era, haber, nosotros, nuestro, también, tras, **vosotros**, **vuestro** |
| Economy | **costo**, dólar, monto, peso, **plata**, productor | empleo, euro, **coste**, financiación, importe, peseta, |
| Education | docente, escuela, estudiante | asignatura, curso |
| Others | *Sports:* cancha, mundial, **plantel**, torneo | *Web:* blog, enlace, online, web, |
| | *Religion:* cruz, dios | *Leisure, personal sphere:* celebrar, competición, concurso, disfrutar, encantar, gustar, hotel, película, plaza, premio, restaurante, usuario, viajero, vídeo |
| | *Law:* constitucional, juez, justicia, legislador, ley, penal tribunal | |
| | *Speech verbs:* agregar, expresar, manifestar, mencionar, señalar, sostener | *Other verbs:* adaptar, añadir, **coger**, conceder, disponer, facilitar, ofertar, ofrecer, proceder, suponer |
| | *Other verbs:* armar, brindar, enfrentar, entregar, fortalecer, generar, ingresar, involucrar, **tomar** | *Other nouns:* actuación, ámbito, aplicación, aportación, **coche**, colaboración, dato, edición, entorno, fase, formación, mejora, número, **ordenador**, página, prever, título, tráfico, verano |
| | *Other words:* área, **auto**, | |
| | capacitación, integrante, mil, miles, nuevamente, promedio, indígena, oportunidad, salud | *Other adjectives:* antiguo, habitual, profesional, relativo, sanitario, siguiente |

By "lexical preferences" we mean those words which do not strictly belong to any specific variety of the language but are preferred by speakers of a particular region. For example, *lograr* ('to get') and *ubicar* ('to locate') are preferred in the Americas, whereas *conseguir* and *situar* are more common verbs for the same meanings in Spain, even though all speakers in both regions know all the words. *Tienda* ('shop') is used in some Latin-American countries, but in others it is much more common to say *comercio, negocio* or *colmado*, depending on the country. The same happens with *acabar* ('to finish'), which is usually substituted by *terminar* in the American esTenTen (as *acabar* has a sexual meaning in some American dialects) or with the American adverb *acá* ('here'); in Spain speakers prefer *aquí*. These few examples indicate that the differences in Spanish dialects that can be traced in the corpus are not given only by lexical or semantic features. It would be perhaps more appropriate to speak about tendencies in the selection of words.

One can of course find words that are only used in America or in Spain, for example, *auto* ('car'), one of the American equivalents of Peninsular *coche,* or the American *costo* ('cost') in contrast to Peninsular *coste.* The pronoun *vos* ('you' singular, of Argentina, Uruguay and other countries) is a clear dialectal trace as well, together with the Peninsular *vosotros* ('you' plural) and *vuestro* ('your' plural), which are not used in America. The verb *haber* in the Peninsular list does not correspond to the lexical verb ('there is/are') but to the auxiliary in compound tenses. In Spain the present perfect is a compound tense *(haber + participle)*, whereas in America it has the same form as the past simple (*canté, cantaste…* instead of *he cantado, has cantado…*).

The keywords also show cultural, political, geographical and other characteristics of the region they represent. In the American sample, we find administrative or territorial divisions typical from American countries, such as *municipalidad* ('council'), which in Spain is called *ayuntamiento. Estado* in the American list represents not only a synonym for 'country' but also the 'state', a key administrative unit in Mexico. *Distrito* ('district') and *federal* ('federal') are parts of the proper name *Distrito Federal,* Mexico's capital city, and *distrito* is also a general equivalent of 'capital' in countries such as Colombia. *Peso* in the American and *euro* and *peseta* in the Spanish are of course the different currencies.[4]

Some words from sports indicate both its importance in America and, at the same time, lexical preferences. *Plantel* in the sense of 'sports team' is not used in Spain (where it is a more general, if less frequent, word for 'staff'.) Words such as *dios* ('god') and *cruz* ('cross') could also indicate the importance of religion in Latin America. Other words, such as *indígena* on the American list, indicate ethnic differences. Finally, some of the differences in both lists can be explained for other non-linguistic reasons. For instance, the word *verano* ('summer'), in the Spanish part, is not frequently used by some American speakers (such as those from Antilles) because they do not have meteorological seasons.

As already noted, national top-level domains (TLDs) were used to identify national varieties. We suspect that, in America, it is not as common as in Spain to use national TLDs, with *.com* more often preferred. This could be why words such as *blog, online* or *web* (anglicisms with the same meaning in Spanish as in English) are present among the Peninsular keywords. The large number of words from politics or administration in the American list could be in part because official sites make up a greater share of that part of the web that uses the national TLD. Words from the world of spare time such as *restaurante* ('restaurant') or *gustar* and *disfrutar* ('to like', 'to enjoy') are included in the Peninsular list. Again, a possible explanation is that American commercial websites frequently use general TLDs like *.com,* not national ones.

We also note problems related to Part-of-Speech tagging errors. We must investigate, for instance, if some of the lemmas reflect ambiguous forms such as *ofertar*, showing the results both for this verb ('to put an offer') and from the noun *oferta*. The form *era* (in the grammatical words of the Peninsular list) is plausibly not the noun 'era' but the verb *ser* ('to be') (though it remains a puzzle why these words were keywords of the one part of the corpus versus the other, with or without errors).

## 4. Conclusion and future work

The esTenTen corpus is a very large corpus of contemporary Spanish, available for all to explore in the Sketch Engine, a corpus query tool offering many functions for analysis. It provides extensive linguistic data and also cultural and social information. For metadata, it has the URL and substrings of the URL. While this is limited, it does support subcorpora for specific countries or areas, and in this paper we have explored distances between nineteen national varieties of Spanish, and examined in detail the key differences between the Peninsular and American subcorpora.

A limitation of this study is that we only looked closely at one national variety (the one from Spain) and, in the qualitative study, did not find the characteristic lexis, and the subcorpus biases, for the other national varieties. This

---------

[4] For a contemporary corpus, the presence of *peseta* may seem odd, since they were replaced by Euros in 2002. The version of the European corpus as used for this paper had since been further cleaned, and most of the occurrences of *peseta* were in the material that was removed. In the latest version of esTenTen as available on the Sketch Engine website, *peseta* is no longer a keyword for the Peninsular part.

is something we hope that we (working with colleagues with expertise in the relevant national varieties) will be able to move on to soon.

Another future line of research could be to add texts created in countries in which Spanish is not the official language, using statistical tools to detect them. It would be good to complement the data that we currently have with Spanish from the United States, a country with around 40 million Spanish native speakers.

## References

Baroni, M., Bernardini, S., Ferraresi, A & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation, 43,* 209–226.

Davies, M. (2002) Un corpus anotado de 100.000.000 palabras del español histórico y moderno. *Procesamiento del Lenguaje Natural 29,* 1–27. URL: http://www.corpusdelespanol.org/x.asp.

Kilgarriff, A. (2009). Simple Maths for Keywords. In M. Mahlberg, V. González-Díaz & C. Smith (Eds.). Proceedings *of the International Conference on Corpus Linguistics*. Liverpool: University of Liverpool. URL: http://ucrel.lancs.ac.uk/publications/cl2009.

Kilgarriff, A. (2012). Getting to know your corpus. *Text, Speech, Dialogue (TSD 2012). Lecture Notes in Computer Science, 7499,* 3–15.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.). *Proceedings of the Eleventh Euralex International Congress.* Lorient: Université de Bretagne-Sud, 105–116.

Moreno Fernández, F. & Otero Roth, J. (2007). *Atlas de la lengua española en el mundo.* Madrid: Fundación Telefónica.

Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)* ELRA. Istanbul, Turkey, 2473–2479.

Pomikalek, J. & Suchomel, V. (2012). Efficient Web Crawling for Large Text Corpora. Proceedings of the 7th Web-as-Corpus Workshop, Lyon, France.

Pomikalek, J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*. PhD thesis. Brno: Masaryk University.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK. URL: http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.html.

Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In M. Baroni & S. Bernardini (Eds.). *WaCky! Working papers on the Web as Corpus.* Gedit, Bologna.

Spanish Royal Academy. Corpus de Referencia del Español Actual (CREA). URL: http://corpus.rae.es/creanet.html.