LAUGHING ONE'S HEAD OFF IN SPANISH SUBTITLES: A CORPUS-BASED STUDY ON DIATOPIC VARIATION AND ITS CONSEQUENCES FOR TRANSLATION

Gloria Corpas Pastor

University of Malaga (Spain) / University of Wolverhampton (United Kingdom)

Abstract

Neutralisation of idiomaticity and diatopic variation are common in translation worldwide. This paper presents a case study of Spanish translation trends for the English idiom *to laugh one's head off*. After a cursory look at the notions of transnational and translational Spanish(es) in Section 2, Section 3 describes the translation strategies deployed in a giga-token parallel subcorpus of Spanish-English subtitles. In Section 4, dictionary and textual equivalents (retrieved from the parallel corpus) are studied against the background of two sets of synonymous idioms in 19 giga-token comparable subcorpora of Spanish national varieties. Corpas Pastor's (2015) corpus-based research protocol is adopted in order to uncover varietal differences, detect diatopic configurations and derive consequences for contrastive studies and translation.

Key words

Spanish varieties, translational Spanish, World Spanish(es), transnational Spanish, neutral Spanish, diatopic variation, idioms, comparable corpora, parallel corpus, translation strategies

1. Introduction

Looking for phraseological information is common practice among translators. When rendering idioms, information is mostly needed to find the appropriate equivalent, but, also, to check usage and diasystemic restrictions. One of the most complex issues in this respect is diatopic variation. English and Spanish are transnational languages that are spoken in several countries around the globe. Cross-variety differences as regards idiomaticity range from the actual choice of phraseological units, to different lexical or grammatical variants, usage preferences and differential distribution. In this respect, translators are severely underequipped as regards information found in dictionaries. While some diatopic marks are generally used to indicate geographical restrictions, not all idioms are clearly identified and very little information is provided about preferences and/or crucial differences that occur when the same idiom is used in various national varieties.

In translation, source language textemes usually turn into target language repertoremes, i.e. established units within the target system. Toury's law of growing standardisation helps explaining why translated texts tend to be more simple, conventional and prototypical than non-translated texts, among other characteristic features. Provided a substantial part of translational Spanish is composed of textual repertoremes, any source textemes are bound to be 'dissolved' into typical ways of expressing in 'standard' Spanish. This means filtering source idiomatic diatopy through the 'neutral, standard sieve'.

This paper delves into the rendering into Spanish of the English idiom *to laugh one's head off.* After a cursory look at the notions of transnational and translational Spanish(es) in Section 2, Section 3 analyses the translation strategies deployed in a giga-token parallel subcorpus of Spanish-English subtitles. In Section 4, dictionary and textual equivalents retrieved from the parallel corpus are studied against the background of two sets of synonymous idioms for 'laughing out loud' in 19 giga-token comparable subcorpora of Spanish national varieties. Corpas Pastor's (2015) corpus-based research protocol will be adopted in order to uncover varietal differences, detect diatopic configurations and derive consequences for contrastive studies and translation, as summarised in Section 5. This is the first study, to the best of our knowledge, investigating the translation of *to laugh one's head off* and also analysing the Spanish equivalent idioms in national and transnational corpora.

2. On the concept of 'World Spanish(es)' and other related notions

Spanish is one of the four originally European languages (together with English, French and Portuguese) which have played a vital role in nation-building, imperial expansion, and globalisation of power structures and economies. Against this background, global Spanish appears to have a symbolic status in the linguo-cultural shaping of the transnational identity upon which *la Comunidad Panhispánica* has been built, as opposed to *the Commonwealth of Nations*, *la Francophonie* and *a Lusofonia* (Del Valle 2011: 288).

While the transnational Hispanic community involves reframing Spanish in the wider context of new forms of economic, political, media and social discourses, [1] those issues are clearly beyond the scope of this paper. By contrast, central to our research are the notions of “World Spanishes”, transnational Spanish, transnational Spanish variants (or transnational Spanishes) and standard Spanish.

By analogy to the well-established English term,[2] in this paper the term *World Spanish(es)* will refer to all Spanish varieties used around the world (cf. Valdés and Geoffrion-Vinci 2012: 598). This concept foregrounds the multiplex nature of the language and is akin to Kachru’s (1985) influential sociolinguistic model of concentric circles. Thus, a similar division could be traced between countries or States where Spanish is a nationwide official language, e.g. Spain, Mexico, Argentina, Puerto Rico, etc. (the Inner Circle); an additional official language, as in Equatorial Guinea (the Outer Circle); or considered traditionally a foreign language, e.g. United States, Brazil, Aruba, Andorra, etc. (the Expanding Circle). Conversely, *transnational Spanish* foregrounds the international and multicultural dimension of the language, as well as the multiplicity of national identities represented by the actual language varieties used in each of the countries where Spanish is an official language, i.e. the *transnational Spanish varieties*.

Even though Spanish varieties are almost entirely mutually intelligible, the transnational settings enhance the social implications of language differences, as well as the complexity of identity formation due to linguistic and cultural issues. Choosing an

---

[1] On language as a source of transnational identity or as a valued commodity, and the persistence of linguistic ideologies linked to nationalism, imperialism and post-colonialism, see Oakes (2011), Wright (2004), del Valle (2007), Niño-Murcia (2015), Mar-Molinero and Paffey (2015), and the papers in Dennison (2013), among others.

[2] The term “World English(es)” denotes the many and variegated transnational varieties of English, including not only American and British English, but such varieties as Indian, Pakistani, Australian, and New Zealand English, as well as the English spoken in various African and Asian countries. For further information, see Kachru (1985, 1992), McArthur (2001), Brutt-Griffler (2002), Kachru et al. (2006), Kirkpatrick (2007) and Seargeant (2012).

‘accent’ plus some relevant linguistic features may result into indexing social classes and assigning certain linguo-cultural identities. According to Niño-Murcia (2015: 742):

> In situations of intense language contact, as in the Americas, where Spanish is in contact with countless indigenous languages and with European languages such as English and Portuguese, people become more aware of language variation. Even when the common denominator is the Spanish language, in a transnational community, speakers of different varieties construct their “difference” around language features.

Mutual intelligibility and linguistic similarity is at the heart of the concept of *global* or *standard Spanish*. Although native speakers of Spanish vary in their use of language, their characteristic linguistic features (pronunciation, grammar, syntax, lexis, phraseology, pragmatics) and registers, the various national variants are similar enough to refer to something called Spanish as if it were a single, monolithic language. In this vein, standard Spanish would be conceived as a linguistic variety that can be considered a correct educated standard for the Spanish language.

This tension between standard Spanish (i.e. prototypical) and transnational Spanish (i.e. pluricentric) also has consequences for the ‘kind’ of Spanish (i.e. transnational variety) used in translation (*translational Spanish*). For commercial reasons, the translation industry has adopted the so-called *neutral Spanish* as opposed to localised Spanish for a specific market (i.e. a diatopic variety). Most worldwide top translation companies offer both language levelling and localisation (into a given Spanish variety or even into *Hispanic neutral Spanish*) as an added-value or as part of their usual services. For illustrative purposes, let us quote from the main site of Trusted Translations:[3]

> Neutral Spanish, commonly referred to as Standard Spanish, Global Spanish or Universal Spanish, is a variation of the Spanish language used to

---

[3] http://www.trustedtranslations.com/spanish-language/translation/neutral-spanish.asp. (Accessed June 21, 2016).

> allow the greatest number of Spanish speakers to understand the message without the use of local terminology and certain verb tenses.
>
> Spanish has evolved over thousands of years across many continents. Hence, many variations and “dialects” of the Spanish language have emerged and are in use across the globe. This presents an interesting situation for anyone that seeks to target all or part of the Spanish-speaking market. It provides an opportunity to tailor a particular translation to target a very specific group, making your communication more effective. However, if your intention is to target a broader audience of Spanish speakers, you will need to use a neutral Spanish that will be accepted and understood by the entire Spanish-speaking population.

In this context, neutral Spanish is to be understood as ‘nationality-neutral’ in terms of distinctive linguistic features, especially accent and idioms restricted geographically to a particular variety. But, at the same time, this sort of ‘neutral Spanish’ – very much sought after in translation – would not be ‘neutral’ at all: it seems to be mostly loaded with socio-cultural positive value and prestige, as opposed to other ‘substandard’, diatopically marked varieties.

In addition, the preference for ‘neutral’ or ‘standard Spanish’ is deeply rooted in ideology and political issues. For example, translational Spanish has been particularly influenced by regulations affecting the film industry. After dubbing became compulsory in Spain in 1941, the variety used was the standard, central Peninsular variety (Bravo 2006: 233). This has continued to be the case even after film censorship (and language control) was abolished in the early ’70s. By then, a concept of core, normalised, standard Spanish was firmly established, and this has continued to be the case, despite the existence of other pluricentric norms. Whereas a limited number of these norms are widely accepted as prestigious, “the most common practice continues to be the levelling of linguistic distinctives in pursuit of a ‘neutral’ Spanish that might be understood throughout the Spanish-speaking world” (Paffey 2012:70).

This fact has been traditionally acknowledged in literary and audiovisual translation, even though language varieties contribute to character profiling and to creating socio-cultural identities: “Investigations so far do in fact suggest that dubbed language varieties are likely to be placed closer to a ‘neutral’, uniform written standard, thus failing to portrait sociolinguistic variation.” (Díaz Cintas 2009: 16). The same tendency can be observed in the translation of fiction into Spanish (cf. Agost Cano 1998, Corpas Pastor 1999, Mayoral Asencio 1999, Bolaños Cuéllar 2004, Hurtado Albir 2007, Sumillera 2008, Ramos Pinto, 2009, Sánchez Galvis 2012, etc.).

Finally, the complex nature of translational Spanish is closely related to the phenomenon of *translationese,* i.e., the characteristic linguistic traits exhibited by translated texts, regardless of their source and target languages. In this vein, translated Spanish (as opposed to non-translated Spanish) is believed to manifest certain universal features, as a consequence of the translation process, as well as distinctive lexico-grammatical and syntactic characteristics. These distinctive features are attributable to widespread translation trends (*universals*),[4] and have been explained by Toury’s (1995: 267) laws of standardisation and interference. In translation, the textual relations in the original or source text (textemes) tend to be replaced by more habitual options offered by the target linguo-cultural system (repertoremes), whereas source text features can also find their way into the target text. Thus, texts translated into Spanish would be simpler and easier to follow (*simplification*), less implicit and ambiguous (*explicitation*), more homogeneous and closer to the standard prototype (*convergence*), more idiosyncratic and ‘typical’ (*normalisation*), exhibit deviation from what is normal in non-translated Spanish (*negative transfer*) or, else, conform to the target language due to the existence of similar features in the source language (*positive transfer*).

---

[4] On universals, see Baker (1995), Laviosa (2002), Mauranen and Kujamäki (2004), Corpas Pastor (2008), Ilisei et al (2010), Kenny (2014).

3. Diatopic variability of idioms in translated Spanish: a case study

In translation, there is a strong tendency to replace units that are diatopically restricted with standard components of the target repertoire. By way of illustration, let us quote Pym's (2008: 314) account of his own experience:

> When I put occasional Australianisms into academic texts, thus creating expressions that are rarer than a blue-arsed fly, they either just disappear in translations or are turned into something absolutely standard (if indeed the copyeditors do not eliminate them first). My textemes are converted into repertoremes, much to my chagrin.

In addition, it has been argued that translated language exhibits simplification and normalisation features with regard to idiomaticity, as demonstrated by Lawick's (2007) analysis of a German-Spanish parallel corpus and Marco's (2009, 2011) studies of the translation into Catalan of English novels, based on the COVALT corpus. No wonder, then, that diatopically-marked idioms tend to be normalised in translation, as described in Leppihalme's (2000) study of the translation into English and Swedish of Finnish novels. In the same line, Corpas Pastor's (2015) study of the translation equivalents retrieved by Linguee established a general tendency towards Peninsular Spanish to the detriment of other national varieties, as well as systematic avoidance of diatopically-restricted collocates.

However, transnational language varieties reflect diatopy in manifold ways. There are idioms that can be easily recognised as belonging to a given variety, because they contain regionalisms. An illustrative example is *medirle el agua a los camotes* ('to plan ahead'; lit. 'measure the water for the sweet potatoes'), whose constituent *camote* is a Mexican word. Other idioms are clearly identified as foreign to a particular variety, as in the case of *a toda madre* ('at full tilt', 'great', lit. 'to whole mother'), again a

Mexican idiom which becomes incomprehensible in Spain, even though all three constituents and the pattern can be found in the Peninsular variety.[5] Moresubtle diatopic indexing is achieved through distributional preferences and other frequency issues. For instance, the act of deferring to take a decision can be conveyed by several collocations in Spanish, mutually comprehensible in all national varieties. But national varieties show a clear tendency to use specific collocations: *posponer una decisión* is usually found in Mexican Spanish, *postergar una decisión* would class as Argentinian Spanish, and *aplazar una decisión* is the preferred option in Peninsular Spanish. Interestingly, the latter is also the most commonly used in translational Spanish (Corpas Pastor, 2015).

In order to describe trends regarding idiomaticity and diatopic variation, a case study on the English idiom *to laugh one's head off* (*inf.* 'to laugh a lot, loudly')[6] and its translations into Spanish is presented in this section. It is based on an innovative and protocolised corpus-based approach that could easily be applied to other idioms, other transnational languages and/or any regional variants of a given language.

3.1. Corpus selection and bitext retrieval

Since idioms are low-frequency items and tend to appear mostly in conversation (especially informal, colloquial registers and slang) or literature, fiction subtitles offer a wide breadth of genres for this purpose. In this study, the translation of *to laugh one's head off* has been analysed in a bilingual parallel subcorpus of Spanish-English film subtitles: the Open Subtitles corpus (2011 version) – *OpenSubtitles2011* (Tiedemann,

---

[5] The construction pattern [A TODO/-A X] is also very productive in Peninsular Spanish: *a todo gas/volume/pulmón/tren/trapo,* etc.; *a toda pastilla/leche/velocidad/máquina,* etc.

[6] *Cambridge Dictionaries Online* <http://dictionary.cambridge.org/es/diccionario/ingles/laugh-your-head-off> (Accessed June 24, 2016).

2009).[7] This 8.31 giga-token multilingual parallel corpus comprises 54 languages and is available through *Sketch Engine*.[8] The size of the English subcorpus is 1.4 billion tokens and the Spanish subcorpora is 624 million tokens[9]. The difference in size is attributable to the fact that not all English film subtitles in the corpus have been translated into Spanish. In any case, the largest bitext is for English-Spanish, with approximately 50 million aligned sentences (Lison and Tiedermannn 2016: 927).

Table 1 illustrates all the occurrences of the idiom in the EN-ES parallel subcorpus of *OpenSubtitles2011*. The number of bitexts (24) in this subcorpus contrasts with the extremely low number of occurrences in the other non-fictional parallel corpora available through Sketch Engine[10] (just one bitext in *Europarl3*, as can be seen in example1). In addition, this also shows that fiction subtitles are prone to containing idioms to the detriment of other non-fictional genres, as said before.

(1)
EN: I would like to say that today Osama Bin Laden must be ***laughing his head off*** because, in my country, instead of arresting terrorists, they concentrate on arresting their captors.
ES: que hoy Osama_Bin_Laden debe de estar ***partiéndose de risa*** porque en mi país, en_vez_de detener a los terroristas, se dedican a detener a sus perseguidores.

The *OpenSubtitles2011* corpus has been automatically compiled by downloading data from the *OpenSubtitles.org* repository of subtitles. It should be noted that, like most web-crawled corpora (cf. section 4.1.), this corpus presents some (pre)processing

---

[7] There is an updated version of this parallel subtitle corpus (*OpenSubtitles2016*) from http://www.opensubtitles.org/. It contains 60 languages and improved processing features, more meta-data, better cleaning-up, etc. However, it has not been made available through Sketch Engine. For further information, see Lison and Tiedermann (2016).

[8] http://www.sketchengine.co.uk. (Subscription required). See Kilgarriff et al. (2014).

[9] https://inventory.clarin.gr/resources/browse/opensubtitles2011/54753ca6524211e583bfaa3fc8d33ad8ae345cc70d7949d7a5fb3f60d35b806d/.

[10] The multilingual parallel macrocorpus OPUS in Sketch Engine contains the following corpora: European Central Bank corpus (ECB), European Medicines Agency documents (EMEA), The European constitution (EUconst), European Parliament Proceedings (v3) (Europarl3), PHP manual corpus (PHP), A parallel corpus of the Balkan languages (SETIMES2), Stockholm Parallel Corpora (SPC), Regeringsförklaringen – Declarations of Government Policy by the Swedish Government 8RF), Belgisch Staatsblad corpus (MBS), OfisPublik, TedTalks, hrenWaC, The Tehran English-Persian subtitle corpus (TEP), KDE4 localisation files (KDE4), KDE manual corpus (KDEdoc), OpenOffice, OpenOffice3, OpenSubtitles2011 – Open Subtitles corpus (2011 version), Tatoeba, Translated UN document s (UN) and Translated UN documents (MultiUN).

shortcomings: (a) language identification and segmentation errors (example 2); (b) de-duplication errors (for instance, bitexts no. 386021057, 386034377 and 386074365 are repeated); (c) misrecognition of characters due to OCR errors, particularly ‘i’, ‘I’ and ‘’l’ in English (see bitexts 9 and 20 in Table 1); and (d) spelling (especially Spanish written accents), grammar and orthotypography errors: see bitexts 2, 3, 4, 18, 24, etc., in Table 1.

(2) < *previous* era ... Tengo 43 años Haré 44 en diciembre Tengo 27 años Ja, ja, ja! ***Women peak*** Las mujeres alcanzan la madurez ***at 40*** a los 40 ***And men at 19*** Y los hombres a los 19 ***I ' member laughing my head off*** Recuerdo que me morí de risa ***When_I read that in a magazine*** cuando lo leí en una revista ***I was 20 at the time*** Tenía 20 años entonces ***Now_I 'm starin ' 40*** Ahora tengo los 40 mirándome ***Right in the face*** directamente a la *next* >[11]

Finally, the converted subtitles are aligned across languages, but not within alternative subtitles for the same language (see bitexts 3-4, 15-16 and 17-18 in Table 1).[12]

1	and on that day… I' m gonna be standing front and center just **laughing my** fuckin **'head off**. God! I' m just messin ' around with my brother.
	Y ese día ... yo me **estaré riendo a carcajada tendida**. ¡ Cielos! Sólo estaba bromeando con mi hermano.
2	He was on the ground, bleeding ... And I **laughed my** fucking **head off** . - He threatened to kill me.
	Estaba en el piso, sangrando ... **Y me morí de la risa**. - Amenazó matarme.
3	You laughed when you heard it at the party. I smiled. I didn' t laugh. You **laughed your** goddamn **head off** ! It was all right. It was a scream. It was very funny, yes.
	Te reíste a el [sic] oírlo en la fiesta. Sonreí, no me reí. ¡ **Te raspaste la** desgraciada **garganta**! Estuvo bien. Estuvo de morirse de la risa. Sí, estuvo muy gracioso.
4	You laughed when you heard it at the party. I smiled. I didn' t laugh. You **laughed your** goddamn **head off** ! It was all right. It was a scream. It was very funny, yes.
	Te reíste a el [sic] oírlo en la fiesta. Sonreí, no me reí. ¡ **Te rompiste la garganta**! Estuvo bien. Estuvo de morirse de la risa. Sí, estuvo muy gracioso.
5	She just laughed when she read the letter. She **laughed her head off** . I refuse to hear anything bad about her. How can she laugh about it?
	Sólo se rió cuando leyó la carta. **Se moría de risa**. Me niego a oír algo malo de ella. ¿ Cómo pudo reírse?
6	You must admit my situation is quite ridiculous. If someone saw us, he **would laugh his head off** . Anyone in my place ...
	Admitirá que es bastante ridícula mi situación. Si alguien nos viera **se partiría de risa**. Cualquiera en mi lugar ...
7	One day, he' d come in, and he' d **be laughing his head off** and totally happy. The next day, he' d come in, and he would be depressed and in tears.
	Un_día venía **riéndose** totalmente feliz. A el día siguiente venía ... ... y estaba deprimido y llorando.
8	He' s waiting at the church ... ... with tears on his face. The bride didn' t show up. Old Man Andrews **is laughing his head off** . Everything exaggerated.
	Está esperando en la iglesia ... con lágrimas en los ojos. La esposa no ha aparecido. El viejo Andrews **se está desternillando**. Todo exagerado.
9	Yes. I' il take care of Multe too. Multe?, Look, Mia. Hey you. (Multe **laughing his head off**) (rattling moan) Ah yes. Ah, ah yes. Keep going.
	Debes cuidar a Multe también. Sí. También cuidaré de Multe. ¿ Multe?, Mira, Mia. Eh, tú. Ah sí. Ah, ah sí. Sigue.
10	That bastard sells dope to school children and he lives in that house. He looks up here and he sees me, and he **laughs his head**

---

[11] Our emphasis.

[12] Intra-lingual alignments are possible, though, in the updated version (*OpenSubtitles2016*).

	off . He throws pebbles up.
	Ese hijoputa vende drogas a los niños y vive en esa casa. Me mira y **se parte de risa**. Me tira piedras.
11	I' ve noticed that when i leave the bar, something will happen. When the guys told me, I **laughed my head off** . Why? The Baby is getting spoilt with us but someday, I' il take her away and have her settle down.
	He notado que cuando dejas el bar, algo puede sucederte. Cuando los chicos me lo dijeron, **me lo tomé a broma**. ¿ Por qué? Baby está mimada con nosotros pero algún día me la llevaré y le haré sentar la cabeza.
12	But I just laughed and begged them to keep on doing it! You ... laughed? I **laughed my head off** ! Otherwise, I wouldn' t be here.
	pero me reí y les supliqué que siguieran. - ¿ Se rió? - **Me partía de risa**. Si no, no estaría aquí.
13	This stuff is hysterical. - Really? - I' **m laughing my head off** at this. The- The getting kicked out of obedience school …
	Esto es para morirse de risa. - ¿ De veras? - **Me estoy desternillando de risa**. Que te echaran de las clases de obediencia ...
14	Women peak at 40 And men at 19 I ' member **laughing my head off** When I read that in a magazine I was 20 at the time
	Las mujeres alcanzan su apogeo a los 40 Y los hombres a los 19 Recuerdo que **me moría de la risa** cuando lo leí en una revista Tenía 20 años entonces
15	What' s the matter with it? I don' t think that' s very nice. Go ahead, **laugh your head off** . I' ve been sitting in that chair since 6: 00 this morning. You sat an hour too long, honey.
	¿ Qué tengo de malo? No creo que eso esté bien. Adelante, **ríase.** Llevo sentada en esa silla desde las 6: 00 de la mañana. Se sentó una_hora de más, cariño.
16	What' s the matter with it? I don' t think that' s very nice. Go ahead, **laugh your head off** . I' ve been sitting in that chair since 6: 00 this morning. You sat an hour too long, honey.
	¿ Es que no le gusta? No le parece bonito. Adelante, **ríase lo que quiera**. He estado en esa silla desde las seis_de_la_mañana. La han dejado espantosa.
17	Yeah, very funny! You scared the hell out of me! Go on**, laugh your head off** ! - Very funny! - Where' s your sense of humor, Marie? Here we are. Okay, now I know why you passed your exams so easily.
	Sí, ríete bien. - Me hiciste cagar de miedo. - No. Adelante, **sigue riéndote**. Recuerda que mi familia se mudó aquí hace seis meses, Marie. Hablan francés peor que yo. Aquí estamos. Y veo por qué aprobaste los exámenes en el primer intento.
18	Yeah, very funny! You scared the hell out of me! Go on, **laugh your head off** ! - Very funny! - Where' s your sense of humor, Marie? Here we are. Okay, now I know why you passed your exams so easily.
	Sí, ríete! ¡ Me diste un susto de el demonio! ¡ Adelante, **muérete de risa**! ¡ Muy gracioso! ¿ Dónde está tu sentido de el humor, Marie? Llegamos. Bien, ahora sé por qué aprobaste tu licenciatura a el primer intento.
19	There ain' t anything crooked about this whole thing. You' **d laugh your head off** if you heard the story. Sure, I' m laughing right now.
	No hay nada sucio en este asunto. Se **troncharía** si conociera la historia. Sí, claro, ya empiezo a **troncharme**.
20	That' s not even worth thinking about. Who? Who? No, you' il **laugh your head off** . Who? The Monk. The Monk? Can he play?
	Ni siquiera vale la pena pensar sobre eso. ¿ Quién? ¿ quién? No, te **cagarías de risa**. ¿ Quién? Monk. ¿ El Monk? ¿ Puede jugar?
21	Let me start by asking you an amusing question. Let me start by asking you one. It' il make you **laugh your head off** . Where' s my money?
	Permítame empezar haciéndole una pregunta divertida. Permítame empezar a mí con una. **Se partirá de risa**. ¿ Dónde está mi dinero?
22	Congratulations! I' m free! Free! Long live the bride and groom! You' **re laughing your head off** . Tell us what' s so funny
	Enhorabuena. Yo soy libre, ¡ libre! ¡ Felicidades! ¡ Vivan los novios! Pero, ¿ **de qué te ríes**? Nos gustaría saber también de qué va la cosa.
23	Everything you say and do is so true and wonderful ... and you make it sound so sacred and holy ... when all the time it' s just a gag with you. You' re just **laughing your head off** at those chumps. You think God' s gonna stand for that?
	Todo lo que dices y haces es auténtico y maravilloso ... y lo haces parecer sagrado y santo ... cuando para ti sólo es un chiste. **Te mueres de la risa** de esos tontos. ¿ Crees que Dios lo tolerará?
24	I ought to kill you, Dutch. You must feel just great, lying there **laughing your head off** at me. Yeah, sure. It was easy to take her away from me, wasn' t it? Anytime you wanted to.
	Debería matarte, Dutch. Tienes que sentirte bien, ahí tirado, **riéndote de mi a carcajadas**. Si, claro. Fue fácil arrebatármela, ¿ verdad? En cualquier momento que quisieras.

Table 1. *Bilingual KWIC for* to laugh one’s head off *(OpenSubtitles2011).*

3.2. Translation strategies and textual equivalents

Translation strategies in Table 1 include some paraphrases through simple units (bitexts no. 15, 17, 19), omissions (cf. bitext no. 9), shifts and compensations (cf. bitexts no. 3, 4), but most bitexts class as clear cases of *équivalence*, i.e. the use of an equivalent target language idiom, as when the segment *He looks up here and he sees me, and he laughs his head off* is translated as *Me mira y se parte de risa* (bitext 10). The Spanish idiom *partirse de risa* is one the translation equivalents usually found for the idiom *to laugh one's head off* in bilingual dictionaries (for instance, in CSD[13]) and it occurs four times in the corpus (bitexts 6, 10, 12 and 21). Other equivalents in bilingual dictionaries are *desternillarse de (la) risa* (OSD, CSD), *reírse a mandíbula batiente* (OSD[14]), *reírse a carcajadas* (CESD[15]) and *troncharse de (la) risa* (CSD).

However, in the corpus there are no instances of *reírse a mandíbula batiente* or *troncharse de (la) risa*. There is just one occurrence of *desternillarse de (la) risa* (bitext 13, plus *[para] morirse de la risa* as emphatic rendering of 'hysterical'). Similarly, there is one occurrence of *reírse a carcajadas* (bitext 24), plus **reírse a carcajada [sic] tendida*: a blend of *reírse a carcajadas* + *llorar a moco tendido* ('cry one's eyes out', 'cry very much') or, else, *reírse a carcajadas* + *largo y tendido* (usually with verbs of talking, 'talk at great length about something', 'have a long chat/talk').

The *OpenSubtitles2011* corpus also contains examples of idiom simplification or shortening through their monolexical verbal constituents, which appear to be also partial equivalents in Spanish: there are four occurrences of *reírse* (bitexts 7, 15, 16 and 22), 2 of *desternillarse* (bitexts 8 and 17) and one of *troncharse* (bitext 19). It should be pointed out that in the case of bitexts 16 and 17 some idiomatic intensification is conveyed through paraphrasis and periphrasis (compensation).

There are two more translation equivalents that are not included in the former list of bilingual dictionaries, but that are usual idioms for 'laughing out loud' in

---

[13] *Collins Spanish-English Dictionary* <http://www.collinsdictionary.com/dictionary/english-spanish>. (Accessed June 24, 2016).
[14] *Oxford Spanish-English Dictionary* <http://www.oxforddictionaries.com/spanish/> (Accessed June 24, 2016).
[15] *Cambridge English-Spanish Dictionary* <http://dictionary.cambridge.org/dictionary/english-spanish/> (Accessed June 24, 2016).

Spanish: *cagarse de [la] risa* (one occurrence, bitext 20) and *morirse de [la] risa*, the most frequent equivalent in the corpus with five occurrences (bitexts 2, 5, 14, 18 and 23). It should be noted that bitext 2 also exemplifies simplification through sanitisation, as the adjective 'fucking' in the source text is ignored, and, therefore, omitted, in the Spanish translation: *And I laughed my* fucking *head off* is rendered as *Y me moría de la risa*.

Other renderings make use of partially equivalent idioms, like *tomarse algo a broma* ('not to take something seriously') in bitext 11, or pseudo-equivalents, like *rasparse la garganta* (bitext 3) and *romperse la garganta* (bitext 4) due to a semantic shift triggered by a verb in context (*to scream*) which is the verbal constituent of other idioms that share the same component (*one's head off*) within the same series: *to scream/shout/cry/etc. one's head off* ('scream, etc. very loud and/or a lot').

None of the translation equivalents in Table 1 contain regionalisms; therefore, they are not overtly marked for diatopic restrictions. This fact does not necessarily mean that all idioms belong to one of the preferred 'neutral, standard' varieties (frequently Peninsular Spanish), nor that they are distributed evenly among all national Spanish variants or within transnational Spanish.

The *Diccionario de locuciones verbales para la enseñanza del español* (DLVEP) includes only verbal idioms restricted to Peninsular Spanish, as stated in the Foreword (Penadés Martínez 2002: 9). The DLVEP lists 12 synonymous verbal idioms denoting 'to laugh very much' (Penadés Martínez 2002: 257). All of them share an identical definition ("reírse mucho"), and are labelled according to pre-CEFR[16] fluency levels and/or degrees of formality: (a) [pre-intermediate/neutral] *reírse las tripas*; (b) [pre-intermediate/informal] *mearse de risa*, *mondarse de risa*, *morirse de risa*, *partirse*

---

[16] CEFR−Common European Framework of Reference for Languages, Teaching, Assessment. <http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf>.

de risa,, reírse las muelas, retorcerse de risa, revolcarse de risa, tirarse de risa and *troncharse de risa*; (c) [intermediate/informal] *desternillarse de risa*; and (d) [proficiency/vulgar] *descojonarse de risa.*

Only three of those ‘laughing out loud’ (LOL) verbal idioms appear in the Spanish translations of Table 1: *desternillarse de risa* (informal, intermediate) and *morirse de risa*, *partirse de risa* (informal, pre-intermediate).

Since the DLVEP is not derived from any corpus, the former classification cannot be taken at face-value, just as indicative of native-speakers’ intuitions. A corpus-based study would probably reveal a completely different picture. For example, it could be that only some (or none) of those 12 synonymous ‘Peninsular’ idioms are actually restricted to the variety spoken/written in Spain. It could also be the case that that some (or a subset) of them rather characterise other non-Peninsular varieties, show distributional differences, etc. Alternatively, it could be argued that some (or all) of the 12 ‘Peninsular’ idioms in the DLVEP belong, in fact, to the common core, i.e. the so-called ‘neutral, transnational’ Spanish, and so forth. In what follows, an innovative corpus-based research protocol will be applied in order to study translation strategies and equivalence from the viewpoint of diatopic restrictions.

4. Spanish equivalent idioms in national and transnational corpora

In this section, the 12 ‘Peninsular’ idioms listed in DLVEP as well as other prospective Spanish equivalents with the common constituent *de risa* will be analysed. On the basis of corpus evidence, it will be argued whether the translation equivalents found in *OpenSubtitles2011* corpus are in line with a particular national variety (or certain varieties) of Spanish, belong to 'neutral' Spanish, or rather reflect translationese.

Several (sub)corpora of the TENTEN family are used for the study (cf. 3.1.): the *esTenTen* (general or standard Spanish), the *esEuTenTen* (European or Peninsular Spanish), the *esAmTenTen* (American or Latin American Spanish) plus the 18 subcorpora of national American varieties included in the latter. Firstly, the 12 synonymous idioms in DLVEP (Set A) are detected and extracted (semi-automatically) from the corpora. Then a two-step comparative analysis of these idioms follows, with a special focus on varietal distributive issues. Finally, in order to get the full picture, a second set of synonymous idioms with the common constituent *de risa* (Set B) is established for all Spanish national varieties, general Spanish and core Spanish idioms denoting 'laughing out loud' (LOL).

4.1. Choice of corpora

Spanish is a fairly well-resourced language in terms of reference corpora available. The *Reference Corpus of Contemporary Spanish* (CREA)[17] was the first fully-fledged corpus to cover Peninsular (or European) and American varieties. While it is still operational, the CREA has been subsumed under the recent *Reference Corpus of 21st Century Spanish* (CORPES XXI),[18] a pan-Spanish general corpus of over 170 million words (1975-2014), which is expected to reach over five billion words in 2018 (Real Academia Española n. d.). The third major Spanish transnational reference corpus

---

[17] http://corpus.rae.es/creanet.html. (Free online access).

[18] http://web.frl.es/CORPES/view/inicioExterno.view. (Free online access).

is the *Corpus del Español: 100 million words, 1200s-1900s* (BYU- Davies 2002-).[19] It is also a pan-Spanish corpus of 100 million words from the 13th to the 20th centuries. At present the BYU-Davies (2002-) is being updated and enlarged considerably: from 20 million words (20th century) to two billion words (21st century) by August 2016.[20]

These are balanced, well-designed corpora that have been carefully cleaned up, lemmatised and annotated morphosyntactically (part-of-speech/PoS tagging). However, none of the three corpora can be deemed maximally representative, since they are either outdated to a certain extent (CREA and BYU-Davies) and/or still under construction (CORPES XXI and BYU-Davies). In addition, they present a series of practical limitations, related to size, access, limited query language, unstable in-built corpus management systems and so forth (Corpas Pastor 2015: 236). In particular, their actual size and coverage of national varieties is relatively small. In the case of BYU-Davies, queries cannot even be filtered on a geographical level.[21] For the aforementioned reasons, these reference corpora prove unsuitable for the study of “World Spanishes”, let alone diatopic restrictions of phraseological units.

By contrast, Web corpora create new possibilities for making comparative analyses of frequency and usage across different national varieties of Spanish. Web corpora are “giga-token corpora created by Web crawling and processing (cleaning up) with new-generation boilerplate removal and de-duplication tools” (Corpas Pastor 2015:

---

[19] http://www.corpusdelespanol.org. The BYU-Davies requires registration and subscription (donation) to have full access.

[20] There is yet another large Spanish corpus: The Corpus of Contemporary Spanish (CEA - Corpus del Español Actual). It contains 540 million words (0.54 billion), which have been lemmatised and PoS tagged. However, it cannot be considered a general, reference corpus as it contains the Spanish components of three multilingual parallel corpora (Europarl, Wikicorpus and Multilingual UN Parallel Text 2000-2009). For more information, see Subirats and Ortega (2012).

[21] The new two-billion-word version of the BYU-Davies will allow filtering of query results according to diatopic varieties.

165). There are several free state-of-the-art Web corpora for Spanish. The ESCOW14[22] is a 3.68 billion/giga-token (GT) sentence shuffle Spanish Web corpus created in 2014 within the COW (Corpora from the Web) project (Schäfer and Bildhauer 2013). ESCOW14 has been crawled with Texrex-neuedimensionen, annotated morphosyntactically with FreeLing and made accessible online using the custom web front-end Colibri2. This corpus allows users to restrict the corpus search to specific strata based on metadata annotations for geographical location ("s_country/city matches"). However, corpus searches are far from intuitive, as they require users to be familiarised with CQP query language, regular expressions and the IMS Open Corpus Workbench (CWB).[23] For instance, complex pattern searches for lemma and PoS tags should be possible via CQP, however the procedure is ill-explained and somewhat counter-intuitive.

The *Araneum Hispanicum* (ARANEUM) is another recent GT corpus from the Aranea family of comparable Web corpora (Benko 2014). This Web corpus has been crawled by SpiderLing and PoS-tagged with open-source, free tools. There are two versions available, according to size: the 1.20 GT *Araneum Hispanicum Maius* and the *Araneum Hispanicum Minus*, a 10% sample of the former (Benko 2015a and b). Both versions are already pre-processed and made available through KonText[24] (ARANEUM_KT) and Sketch Engine (ARANEUM_SE). Technically speaking, the ARANEUM allows subcorpus building of national varieties by filtering through Internet country code top/second-level and web domain (e.g., .es, .hn, .mx, .ar, .uy, .ve, etc.). Through Sketch Engine the process is lengthy and time-consuming, as URLs have

---

[22] http://corporafromtheweb.org/escow14/. (Free online access for academics only; registration required). ESCOW14 replaces ESCOW12, created in 2011 with a former version of the processing software (Texrex-mrvain) and TreeTagger (Schäfer and Bildhauer 2012).

[23] For further information, see http://cwb.sourceforge.net/files/CQP_Tutorial/.

[24] https://www.kontext.korpus.cz.

to be selected manually. By contrast, automatic diatopic selection is possible with KonText, but then query types are extremely limited.

Finally, the *esTenTen* (TENTEN) is a giga-token Web corpus of general Spanish created automatically in 2011 (Kilgarriff and Renau 2013). This modular macrocorpus is meant to be representative of the global, standard language spoken/written across the Spanish-speaking world. It is composed of two large corpora: the 2.3 GT *esEuTenTen* corpus of European Spanish (Peninsular variety) and the 8.6 GT *esAmTenTen* corpus of Latin American Spanish (American variety).[25] Corpora have been crawled with Spiderling[26], tokenised, lemmatised and PoS tagged with Freeling, and made available through Sketch Engine[27]. National varieties have been identified by their national top-level domains (TLDs: .ar, .bo, .cl, .co, .es, .uy, etc.),[28] which allows users to restrict searches by just one particular variety, or a subset of varieties. In addition, all national subcorpora contain roughly "the same mix of different text types", as they have been compiled "using exactly the same method" (Kilgarriff and Renau 2013: 14).

It should be noted that the three national varieties with the largest corpus size correspond to Argentina, Spain and Mexico; and that 75% of the American Spanish corpus comes from Argentina, Mexico and Chile, in descending order (Suchomel and Pomikálek 2012). The actual size (million words) of the national varieties in the *esTenTen* corpus can be found in Table 2:

COUNTRY	TLD	SIZE
Argentina	**.ar**	2,447
Bolivia	**.bo**	47
Chile	**.cl**	859

---

[25] Similarly to the way linguists on both sides of the Atlantic often speak of "Americanisms" vs. "Peninsularisms" (cf. Lipski 2012: 19), and following Kilgarriff and Renau (2013), in this paper we will use the term "Peninsular Spanish" to refer to the Spanish variety spoken/written in Spain; and "American Spanish" as an umbrella term to refer to the American variety represented by the 18 Latin American national varieties included in the *esAmTenTen corpus*.

[26] http://nlp.fi.muni.cz/trac/spiderling. See also Pomikalek and Suchomel (2012).

[27] https://www.sketchengine.co.uk/documentation/wiki/SkE/Help/JargonBuster.

[28] The *esAmTenTen* corpus does not cover the Spanish varieties spoken/written in Puerto Rico or southwestern United States.

Colombia	**.co**	371
Costa Rica	**.cr**	47
Cuba	**.cu**	211
Dominican Republic	**.do**	43
Ecuador	**.ec**	64
El Salvador	**.sv**	27
Guatemala	**.gt**	27
Honduras	**.hn**	8
Mexico	**.mx**	1,470
Nicaragua	**.ni**	53
Panama	**.pa**	15
Paraguay	**.py**	51
Peru	**.pe**	253
Spain	**.es**	1,992
Uruguay	**.uy**	156
Venezuela	**.ve**	218

Table 2. *National varieties in* esTenTen *(Kilgarriff and Renau 2013: 13-14)*

For this paper the TENTEN macrocorpus (general, standard Spanish), corpora (Peninsular and American varieties) and subcorpora (Latin American national varieties) have been selected for various reasons. On the one hand, the Spanish TENTEN is modular: thanks to the way the Peninsular and American components have been crawled, assembled as subcorpora and computed for similarity, diatopic filtering can be performed easily, fast and accurately through TLDs. On the other hand, the TENTEN macrocorpus, corpora and subcorpora are already processed and made accessible through Sketch Engine, a robust, full-fledged and intuitive query system (QS). Thus, the (sub)corpora selected can be exploited to their full potential through all core functions and other functionalities offered by Sketch Engine, including lexical and grammar patterns through WordSketch. This makes TENTEN more versatile for our research purposes than other Web-crawled Spanish corpora (see Table 3). And, finally, the results of this study could be easily compared with those previously obtained by Corpas Pastor (2015) on geographical variations of Spanish collocations, also based on the TENTEN corpora.

FEATURES	CORPORA			
	ESCOW14	**ARANEUM_KT**	**ARANEUM_SE**	**TENTEN**
Diatopic filtering	√	√	*	√
Intuitive QS	–	√	√	√
Lemma Search	√	√	√	√
Phrase search	√	√	√	√
Frequency list	*	–	√	√
N-grams	*	–	√	√
Patterns	*	–	√	√

Table 3. *Usability features for accessing Spanish Web Corpora*[29]

4.2. Semi-automatic detection of idioms

The extraction techniques and association measures currently used for contextual idiom detection, such as mutual information, metric clusters, corpus frequency ratio, etc., tend to retrieve better results for collocations than for idioms (cf. Heid 2008, Corpas Pastor 2013). The main difficulties when it comes to (semi-)automatically detecting idioms are attributable to their own nature (polylexicality, ambiguity, discontinuity, free slots, frozenness and variability, etc.); to their creative uses in discourse; to the size and quality of the analysis corpus (idioms tend to be low-frequency items); and to the performance of a specific NLP approach.[30] In the case of giga-token size corpora crawled from the Web, there are additional problems which have to do with corpus preparation and processing (document selection and cleaning up) and with the parsing and annotation systems in place. Wrong part-of-speech (PoS) tagging or parsing, as well as grammar and punctuation errors, substandard spellings or typos could result in low precision (noise) and recall (silences), i.e. non-idioms being identified/quantified as idioms and n-grams (literal/unrelated) being wrongly retrieved as idiom candidates; or, else, idioms not being identified/quantified as idioms.

---

[29] √ (feature present and user-friendly), * (feature present, but cumbersome or difficult to use), − (feature not present).

[30] See also Sag et al. (2002), Heid (2008), Corpas Pastor (2013, 2015), Corpas Pastor et al. (2013), Feldman and Peng (2013), among others. For further reference, visit the SIGLEX-MWE website (http://multiword.sourceforge.net/PHITE.php?sitesig=MWE).

On a note of caution, it should be pointed out that the concordances in the TENTEN (sub)corpora may reflect processing and other PoS/parsing errors, which could compromise results. For instance, with a window of five (left and right), the lemma search sequence [(partir)-v + (risa)] yields example 3 due to the wrong PoS and lempos (PoS suffix conjoined to a lemma) assigned to *partir* (verb) in the complex preposition *a partir de* (+ noun phrase). Examples 4-5 evidence parsing errors for search sequences [(retorcer)-v + (risa)] and [(revolcar)-v + (risa)]. Example 6 illustrates substandard spelling in combination with wrong PoS tagging for the search sequence [(mear)-v + (risa)]: the personal pronoun *me* (misspelled *mee*) is wrongly assigned to the verbal lemma *mear* (a first person singular past tense form). Finally, typos and shortenings or abbreviations in the corpus can prevent the detection of idioms, as seen in example 7 (*revolcarse de risa*) and 8 (*mearse de risa*):

(3) cosecha artística de la ***risa*** en Cuba , a ***partir*** de esta escuela que vertía sus enseñanzas
(4) carnosa y blanca; una ***risa*** horrorosa parecía ***retorcer*** sus rasgos en una mueca eterna
(5) creía que iba a reventar de la ***risa***. Se ***revolcaba*** por el suelo, se apretaba la barriga, se
(6) rato q no andaba x aca, muy bueno todo y ***mee*** caguee de ***risa*** con lo de la vieja estación
(7) de atras... la hilux se esta ***rebolcando*** de ***risa*** con esta BASOFIA, espero que la vw amarok
(8) ??? jajaja perdon pero aqui si me ***meo*** d ***risa***, eres de los que creen que por pinches 4

Better results can be obtained when the idiom kernel [DE RISA] is used as node, especially in the presence of lexico-syntactic and morphological flexibility (Corpas Pastor 1995, 2013). An idiom kernel is the minimum core of constituent(s) needed to recall/identify a given idiom or sub-group of idioms sharing the same kernel constituent(s). Thus, a query search of phrase/kernel plus context (positive filter) like [(de, risa)] + (partir)] retrieves more accurate results (e.g. 9-11), although some noise remains (12-14).

(9) flipandooo en colorines y me empece ***a partir de risa*** ... (pero como todo el mundo) y encima me
(10) algo de tripita de los partos y me ***parto de risa*** cuando estiro en el gim.manos hacía arriba
(11) su absurdo guión. Habrá quien se ***parta de*** la ***risa*** con su humor profundamente japonés y habrá
(12) los estimulaba físicamente, los tentaba ***de risa***, para que a ***partir*** de esa pérdida del
(13) absolutamente original, a ***partir*** de clases llenas ***de risa***, ritmos, historias, colores y sabores!
(14) prontos a ***partir*** a un inolvidable viaje ***de risa*** y buen humor. Una comedia de primera clase

Other general problems experienced when extracting idioms in the TENTEN corpora are related to language-detection and de-duplication. In the first case, the corpora contain some English words and sentences (example 15), as well as sentences in other romance languages that have not been automatically detected and removed (example 16). In the second case, there is repetition of sentences, due to the fact that de-duplication with Onion (ONe Instance ONly)[31] is performed at paragraph level, not at sentence-level, which is considered to be too small a unit: e.g., example 17 appears four times in the *esTenTen* corpus with four different ID numbers (#1628794172, #1628794244, #1628794163, #1628794235):

(15) powerful works of art. And you finally have the chance to catch a glimpse. India (Hindi :
(16) pegaba un brinco e o home veña a pór carade risa. Hai que ser ....!! <gap tokens_count="31" />
(17) entiende mi miedo y mi reaccion y se cgan de risa!!! y yo a punto de quedar dura de los nervios

4.3. Results and discussion

The 12 synonymous LOL idioms listed in DLVEP (henceforth Set A) are typical of the national variety spoken/written in Spain. Based on this observation, Set A idioms would be expected to appear mainly in the Peninsular variety. However, it could be the case that all/some of them (i) are actually restricted to Spain, (ii) are indicative of other

---

[31] On Onion, see Pomikálek (2011).

national varieties, or, else, (iii) belong to core, general Spanish. It could also be possible that there are other instances of synonymous idioms (Set B) in the corpora analysed which could fall under one or all three former cases.

4.3.1. LOL idioms (Set A)

Set A idioms have been detected (semi-)automatically and their frequencies computed in (a) the *esEuTenTen* corpus of Peninsular Spanish (PenSp), (b) the *esAmTenTen* corpus of Latin American Spanish (AmSp), and (c) the *esTenTen* corpus of general Spanish (genSp). The results are summarised in Table 4. Columns 2-4 show raw/normalised frequencies for each idiom in column 1. Raw frequencies (first row) refer to the total number of occurrences (actual instances) of a given idiom in the corpus, whereas normalised frequencies (second row, in bold) are provided as percentage scores. Normalising frequencies to a common base (e.g., per million tokens) is needed in order to compare results in corpora of different sizes. For instance, there are 20 instances of *retorcerse de risa* in PenSp, but seven times more in AmSp (140 instances) and six more in genSp (120 instances). However, this does not mean that this idiom is more frequent in Latin America than in Spain, as they have the same normalised frequency (0.01) in both varieties as well as in standard general Spanish. And vice versa, similar frequency counts do not necessarily imply similar frequency of occurrence per million words: while *mondarse de risa* appears 42 times in PenSp and 43 in genSp, their normalised frequencies differ only slightly (0.02 vs. 0.00). Thus, in what follows, only normalised frequencies will be taken into account in order to compare data distribution.

VERBAL IDIOMS	esEuTenTen	esAmTenTen	esTenTen
descojonarse de risa	35	3	16
	0.01	**0.00**	**0.00**
desternillarse de risa	73	238	310
	0.03	**0.03**	**0.03**
mearse de risa	255	284	539
	0.11	**0.03**	**0.05**
mondarse de risa	42	3	45
	0.02	**0.00**	**0.00**
morirse de risa	1,377	4,840	6,222
	0.58	**0.56**	**0.57**
partirse de risa	1,538	525	2,063
	0.65	**0.06**	**0.19**
reírse las muelas	–	–	–
	–	–	–
reírse las tripas	3	–	3
	0.00	**–**	**0.00**
retorcerse de risa	20	120	140
	0.01	**0.01**	**0.01**
revolcarse de risa	13	119	132
	0.01	**0.01**	**0.01**
tirarse de risa	67	20	80
	0.03	**0.00**	**0.01**
troncharse de risa	140	22	118
	0.06	**0.00**	**0.01**

Table 4. *Raw and normalised frequencies (PenSp, AmSp & genSp).*

The first thing that springs to mind is that most idioms of ‘laughing’ analysed in column 1 are to be found in all three corpora, with the exception of *reírse las tripas* (only in PenSp, rare: 3/0.00) and *reírse las muelas* (not found in any corpus). Therefore, both idioms have been removed from the list of LOL idioms in Set A. The ten remaining idioms are distributed along a frequency rank (FR): Rank I (0.50-0.56), Rank II (0.19-0.11), Rank III (0.06-0.03) and Rank IV (0.01-0.00). The first two positions in the frequency list (FL) are occupied by the same two idioms (Ranks I-II): *morirse de risa* and *partirse de risa* (more frequent in PenSp). The third position is also occupied by the same idiom in the three corpora (Ranks II-III): *mearse de risa*. Position number 4 is occupied by the idiom *desternillarse de risa* in genSp and AmSp, while in PenSp appears in fifth position (see Table 4).[32]

---

[32] Idioms are listed by descending order according to frequency rank (I-IV, shadowed differently) and position (standardised/raw frequencies) within each rank (1-10).

FL	PenSp	AmSp	genSp
1	*partirse de risa*	*morirse de risa*	*morirse de risa*
2	*morirse de risa*	*partirse de risa*	*partirse de risa*
3	*mearse de risa*	*mearse de risa*	*mearse de risa*
4	*troncharse de risa*	*desternillarse de risa*	*desternillarse de risa*
5	*desternillarse de risa*	*retorcerse de risa*	*retorcerse de risa*
6	*tirarse de risa*	*revolcarse de risa*	*revolcarse de risa*
7	*mondarse de risa*	*troncharse de risa*	*troncharse de risa*
8	*descojonarse de risa*	*tirarse de risa*	*tirarse de risa*
9	*retorcerse de risa*	*descojonarse de risa*	*mondarse de risa*
10	*revolcarse de risa*	*mondarse de risa*	*descojonarse de risa*

Table 5. *Frequency ranks for Set A idioms (PenSp, AmSp & genSp)*

However, the remaining idioms show a considerable degree of variation as regards their frequency ranks and normalised scores in the two main geographical (e.g. Continental)[33] variants. The Latin American variety appears to be closer to general Spanish (80% total coincidence)[34] than Peninsular Spanish (20% total coincidence). American and Peninsular Spanish show only 10% total coincidence, although percentages rise when it comes to idioms within the same rank (30%) or within the first five positions of the frequency list (40%). Partial coincidence among Peninsular Spanish and general Spanish is also higher as regards idioms belonging to the same rank (50%), but decreases to 10% when comparing the actual number of idioms in the same position of the frequency list (just one, *mearse de risa*).

So far idiom distribution has been established for the two main Continental varieties among themselves (PenSp and AmSp) and in relation to the corpus of general Spanish. In Corpas Pastor (2015), diatopic restrictions of collocations have also been found in both main Continental varieties, and for each national variety of Spanish (including Peninsular Spanish), as well as diatopic preferences in the choice of verbs for particular

---

[33] In this paper the term “Continental varieties” is used to refers to Peninsular Spanish (in Europe) as opposed to general Latin American Spanish (in America).

[34] Total coincidence implies the same rank and position within the list. Partial coincidence applies to idioms within the same rank or in the same position in the list (FR).

semantic and functional values. Focusing on the three national varieties with the largest corpus size in the TENTEN corpora (Argentina, Mexico and Spain), it was established that (a) almost 70% of all verbal collocates in the three main national varieties coincide (at least partially) with general Spanish; (b) over 30% of significant verbal collocates for the type *V.* + decisión_n in the three national varieties were diatopically restricted; (c) the national varieties closer to general Spanish are Mexican Spanish (60% shared verbal collocates), followed by Argentinian and Peninsular Spanish (only 45% shared verbal collocates); and (d) the closer national varieties among themselves are Mexico-Spain (32.43%), followed by Mexico-Argentina (29.72%) and Argentina-Spain (24.32%), as the most distant national varieties.

Against this background, the so-called "Latin American Spanish" looks, in fact, an amalgam of national varieties, where Mexican seems to occupy a central position, both as regards Latin American varieties and Peninsular Spanish. Therefore, a second level of analysis is needed in order to ascertain relevant varietal differences among the 19 national varieties of Spanish (including Peninsular Spanish). Table 6 summarises the main findings as regards the ten idioms selected (Set A).

IDIOMS	.AR	.BO	.CL	.CO	.CR	.CU	.DO	.EC	.SV	.GT	.HN	.MX	.NI	.PA	.PY	.PE	.UY	.VE	.ES
descojonarse de risa	2	–	2	1	–	–	–	–	–	–	–	4	–	–	–	1	1	1	35
	0.00	**–**	**0.00**	**0.00**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**0.00**	**–**	**–**	**–**	**0.00**	**0.00**	**0.00**	**0.01**
desternillarse de risa	104	1	20	16	–	–	–	–	2	–	–	36	1	1	1	16	11	8	73
	0.03	**0.00**	**0.02**	**0.03**	**–**	**–**	**–**	**–**	**0.07**	**–**	**–**	**0.02**	**0.01**	**0.05**	**0.01**	**0.05**	**0.05**	**0.03**	**0.03**
mearse de risa	2	–	49	3	–	–	–	–	–	–	–	37	–	–	–	9	1	–	255
	0.00	**–**	**0.04**	**0.01**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**0.02**	**–**	**–**	**–**	**0.03**	**0.00**	**–**	**0.11**
mondarse de risa	1	–	–	–	–	–	–	–	–	–	–	1	–	–	–	–	–	–	42
	0.00	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**0.00**	**–**	**–**	**–**	**–**	**–**	**–**	**0.02**
morirse de risa	1,904	3	1,014	179	5	23	9	1	5	8	1	963	33	24	5	247	93	68	1,377
	0.57	**0.05**	**0.87**	**0.36**	**0.10**	**0.10**	**0.13**	**0.01**	**0.18**	**0.29**	**0.12**	**0.49**	**0.47**	**1.14**	**0.07**	**0.73**	**0.44**	**0.23**	**0.58**
partirse de risa	220	1	70	13	–	7	13	1	1	1	–	104	4	2	–	40	9	2*	1,538
	0.07	**0.00**	**0.06**	**0.03**	**–**	**0.03**	**0.30**	**0.01**	**0.03**	**0.03**	**–**	**0.05**	**0.06**	**0.10**	**–**	**0.12**	**0.04**	**0.00**	**0.65**
retorcerse de risa	62	–	15	4	–	–	–	–	–	–	–	29	1	–	1	6	–	–	20
	0.02	**–**	**0.01**	**0.01**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**0.01**	**0.01**	**–**	**0.01**	**0.02**	**–**	**–**	**0.01**
revolcarse de risa	47	–	16	4	–	–	–	–	–	–	–	32	1	–	–	15	2	–	13
	0.01	**–**	**0.01**	**0.01**	**–**	**–**	**–**	1	**–**	**–**	**–**	**0.02**	**0.01**	**–**	**–**	**0.04**	**0.00**	**–**	**0.01**
tirarse de risa	132	–	–	–	–	–	–	**0.01**	–	–	–	42	–	–	–	3	–	–	67
	0.04	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**0.02**	**–**	**–**	**–**	**0.01**	**–**	**–**	**0.03**
troncharse de risa	17	–	–	–	–	–	–	–	–	–	–	6	–	–	–	1	–	–	140
	0.01	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**0.00**	**–**	**–**	**–**	**0.0**	**–**	**–**	**0.06**

Table 6. *Raw and normalised frequencies of idioms in the 19 national varieties of* esTenTen *(Set A).*

According to Lipski (2012: 3), there are ten dialectal areas in Latin America (or 'American dialects': 1. Mexico (except for coastal areas) and southwestern United States; 2. the Caribbean region (Cuba, Puerto Rico, Dominican Republic, Panama, Caribbean coast of Colombia and Venezuela, and Mexico's Caribbean and Pacific coasts); 3. Guatemala, parts of Mexico's Yucatan Peninsula, and Costa Rica; 4. El Salvador, Honduras, and Nicaragua; 5. Colombia (interior) and neighbouring highland areas of Venezuela; 6. the Pacific coast of Colombia, Ecuador and Peru; 7. Andean regions of Ecuador, Peru, Bolivia, northwest Argentina, and northeast Chile; 8. Chile; 9. Paraguay, northeastern Argentina, and eastern Bolivia; and 10. Argentina (except for the extreme northwest and northeast) and Uruguay.

In this respect, the results of our analysis prove to be quite revealing. According to corpus data, all 18 varieties (100%) appear to use the idiom *morirse de risa* (albeit with different frequencies), while *partirse de risa* can be found in 15 varieties (83.33%). The average number of different idioms per national variety is 4.5. However, the distribution of idioms types in the national subcorpora seems to draw a dividing line which separates most Caribbean and Central American countries (with notable exceptions) from South American countries. Thus, there is very little presence of those idioms in countries belonging primarily to dialectal areas 2-4 (1-3 idioms): Costa Rica, Cuba, Dominican Republic, El Salvador, Guatemala, Honduras and Nicaragua, with the exception of and Panama (four idioms); dialectal areas 6-7 (Ecuador, 3 idioms). For all those national varieties the mean values are below the average (4.5 idioms). The richer dialectal areas (in terms of number of different idioms from Set A) are to be found towards the Northern and Southern parts of Latin America: Mexico (ten idioms, dialectal areas 1-3) and Argentina (ten idioms, dialectal areas 7, 9-10). The subcorpora of national varieties for most countries situated along the western part of the Andes mountain range show a number of idioms higher than the average:

Colombia (7), Peru (9), Chile (7), whereas neighbouring countries situated towards the east tend to register values lower than average: Venezuela (4), Bolivia (3) and Paraguay (3), with the exception of Uruguay (6), possibly because of Argentina's geographical proximity and influence.

Another interesting finding is the distribution of idiom tokens (occurrences/instances) in general Spanish and per national varieties (see Tables 3 and 6). The Mexican and Argentinian national varieties appear equidistant from general Spanish as regards idioms' normalised frequencies (four coinciding values out of ten idioms: 40% coincidence), whereas Peninsular Spanish appears slightly more distant (three coinciding values, 30%). In any case, these results appear somewhat inconclusive, as coinciding values are actually 0.00/0.01. The same can be observed in terms of the actual distance among the three national varieties: Argentina-Mexico (20%), Argentina-Spain (20%), Mexico-Spain (10%), with values of 0.00/0.01, except for *desternillarse de risa* (0.03 in Argentina and Spain). The resulting picture differs from the situation of diatopically restricted collocations described in Corpas Pastor (2015). One possible explanation could be that idioms are less frequent in corpora than collocations, which makes it more difficult to observe diatopic variation.

On the other hand, intervarietal differences seem to be more outstanding the higher the normalised frequency of a given idiom. Take, for instance, the two most frequent idioms from the list in general Spanish: *morirse de risa* (0.46) and *partirse de risa* (0.10). In the case of the former, differences range from + 0.68 (Paraguay) to – 0.41 (Bolivia); in the latter case, differences can reach up to + 0.55 (Spain), with -0.10 as the lowest value (Bolivia and Venezuela). By contrast, there is little intervarietal difference (between -0.01 and + 0.03) for very low-frequency idioms, like *descojonarse de risa* (0.00), *retorcerse de risa* (0.01) and *revolcarse de risa* (0.01).

Furthermore, diatopic indicators could be developed by means of divergent configurations (vectors of features or conglomerates) that take into account the actual presence of the idiom and its average frequency in a given variety or set of varieties. For instance, a strong presence of *morirse de risa* combined with the absence of *troncharse de risa* would clearly indicate Paraguay and Chile; average values for *morirse de risa* and high values for *troncharse de risa* would be indicative of Peninsular Spanish; whereas a value of 0.03 for *desternillarse de risa* in combination with values below average for *morirse de risa* would be indicative of Colombia and Venezuela, and so forth (cf. also Tables 3-4).

4.3.2. LOL idioms (Set B)

It should be borne in mind that the results in 2.3.1. have been obtained for a pre-defined list of synonymous idioms identified as typical of Peninsular Spanish in DLVEP (Set A). Results would probably vary if the corpora selected for this study were freely searched for other synonymous idioms. With this aim, a third analysis has been carried out for LOL idioms different from Set A and with the same pattern: a reflexive verb (or verb with a reflexive pronoun) and the same non-verbal constituent. Phrase searches have been performed with the idiom kernel [DE RISA] and left-sorted PoS filter V. (= verb). Once concordances had been retrieved, KWIC lines were examined to discard non-synonymous sequences (example 18), synonymous idioms composed of a reflexive verb and kernel plus other constituents (example 19), with the kernel and another constituent as subject (example 20), or with non-reflexive verbal components (example 21), and hapax legomena (example 22).

(18) claro, que lo de la automatrícula ***ha sido de risa*** para desgracia nuestra. Es un hecho que
(19) medios europeos ***se están partiendo el culo de risa*** con los españoles y las imágenes de
(20) tiempos y a uno ***se le saltan las lágrimas de risa*** pensando en aquellas generaciones de

(21) tal modo las tripas, que el juez ***llorando de risa*** le acuso de desacato. El sonrojo le duró (2
(22) además de interesante y demostrativo, para ***wisharse de risa***. Juan E preparó un diálogo

Filtered results show a large number of verbal constituents (over 60) associated to the idiom kernel [DE RISA] in the sense of ‘laughing out loud’. Table 7 lists national varieties in descending order as regards number of verb types found and corresponding percentage from the total number of Spanish verb types (67). For ease of comparison, subcorpora sizes are provided again in the fourth column. More than 50% of all verbal variants are found in the Spanish of Spain, followed by that of Argentina (34.32%), Peru (32.83%) and Mexico (26.85%). Less than 10% of the Spanish verbal variants available are used in those four national varieties (Spain, Argentina, Peru and Mexico); over 5% in Colombia, Chile, Uruguay and Cuba; and between 4% and 0% in Costa Rica, Dominican Republic, El Salvador, Bolivia, Ecuador, Guatemala, Nicaragua, Paraguay, Venezuela, Panama and Honduras. Although the size of the national subcorpora may be playing a role in idiom verbal variability, it does not seem to be the only determinant factor. While it is true that the Spanish national variants with larger corpus sizes tend to have more idiom verbal variants (e.g., Spain, Argentina and Mexico), and vice versa (the smaller the size, the fewer variants, e.g. Panama and Honduras), a clear correlation cannot established for all cases. For instance, a smaller-sized subcorpus like Peruvian Spanish (253 million words) contains the third higher percentage of verbal variability (32.83%), just below Argentinian (35.82%) and Mexican varieties, whose corpora are at least six times larger. By contrast, the Chilean subcorpus, more than three times bigger than the Peruvian, contains only eight variants (11.94%), whereas Venezuelan Spanish, with a similar-sized subcorpus (210 million words) registers only one out of 67 idiom verbal variants (1.49%).

NATIONAL VARIETY	NO.	%	SIZE
Spain	34	50.74	1,992
Argentina	24	35.82	2,447
Peru	22	32.83	253
Mexico	18	26.86	1,470
Colombia	9	13.43	371
Chile	8	11.94	859
Uruguay	8	11.94	156
Cuba	6	8.95	211
Costa Rica	4	5.97	47
Dominican Republic	3	4.47	43
El Salvador	3	4.47	27
Bolivia	2	2.98	47
Ecuador	2	2.98	64
Guatemala	2	2.98	27
Nicaragua	2	2.98	53
Paraguay	2	2.98	51
Venezuela	1	1.49	218
Panama	1	1.49	15
Honduras	0	0.00	8

Table 7. *Verbal constituents per Spanish national variety (Set B).*

Results based on the TENTEN corpora are not to be taken at face value due to the limitations already mentioned in Section 4.2., namely processing and parsing errors, substandard orthotypography, typos, shortenings, abbreviations, incomplete deduplication or cleaning up, etc. By way of illustration, the most widely distributed idiom in Set B (*cagarse de risa*) appears rather difficult to identify in the subcorpora due the wide range of substandard spellings found for its verbal component (see examples 23-27). Therefore, it is not possible to detect all actual instances of this idiom in the various subcorpora for obvious reasons.

(23) ridiculo total, pero nos vamos a ***c********   ***de risa*** con las salidas de este plancha, pobres

(24) colores de las bolsas estan para ***cag@#$ de risa*** ... espero que Ladislao Kubala publique

(25) </p><p>Yo estoy en el foro porque me ***kgo de risa*** al leer filosofos (i4everluis), economistas

(26) financie, cantin ja, sabes como nos ***c-g-m-s s de risa*** con tus virtudes artisticas, asi necesita

(27) adentro. Simple. Andrés: me hiciste ***C464R de risa***! VeroS: podés votar en blanco para

However, the results obtained could be deemed sufficiently illustrative of the Spanish present situation in general. The distribution of Set B idioms in the 19 Spanish national variants is provided in Table 8. None of those verbal constituents is used in all varieties. The most generalised idioms are *cagarse de risa* (found in 10 national variants), *matarse de risa* (8), *doblarse de risa* (8) and *destornillarse*[35] *de risa* (7). But most idioms are restricted to a handful of national variants, or their use is unmistakably marked. For example, the following set of verb components for V_{refl}_*de risa: cascarse, crujirse, descacharrarse, descoyuntarse, desgüevarse, desmontarse, despelotarse*, among others, seem to be restricted diatopically to Peninsular Spanish; *atacarse, cargarse, pavonearse* and *zurrarse* + *de risa* can only be found in the Mexican subcorpus; *asfixiarse, atorarse, desintegrarse* and *desentornillarse* + *de risa* could be considered Peruvianisms; *desencajarse, despanzarse, desparramarse* and *pujarse* + *de risa* also appear as Argentinisms; *miarse de risa* is only documented in the Uruguayan subcorpus, and so forth.

---

[35] The idiom *destornillarse de risa* may have originated by folk etymology from the quasi-homonymous idiom in Set A (*desternillarse de risa*). Both variants co-exist in Argentina, Chile, Mexico, Spain and Uruguay. Interestingly enough, *destornillarse de risa* is the only variant documented in Costa Rica and Peru (also *desentornillarse*). The idiom *desternillarse de risa* is the only variant available in Bolivia, Colombia, Nicaragua, Panama, Paraguay and Venezuela.

V_{REFLX}_DE RISA	.AR	.BO	.CL	.CO	.CR	.CU	.DO	.EC	.SV	.GT	.HN	.MX	.NI	.PA	.PY	.PE	.UY	.VE	.ES
ahogarse	x	−	x	x	−	x	−	−	−	−	−	x	−	−	−	x	−	−	−
arrastrarse	−	−	−	−	−	−	x	−	x	−	−	−	−	−	−	x	−	−	−
asfixiarse	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x	−	−	−
atacarse	−	−	−	−	−	−	−	−	−	−	−	x	−	−	−	−	−	−	−
atascarse	−	−	−	−	−	−	−	−	−	−	−	x	x	−	−	−	−	−	−
atorarse	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x	−	−	−
atracarse	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x	−	−	−
atragantarse	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x	−	−	x
botarse	−	−	−	−	−	−	−	−	−	−	−	x	−	−	−	−	−	x	−
caerse	−	−	−	x	−	−	x	−	−	−	−	x	−	x	−	−	−	−	x
cagarse	x	x	x	x	x	x	−	x	x	x	−	x	x	−	x	x	x	−	x
carcajearse	−	−	−	−	−	−	−	−	−	−	−	x	−	−	−	x	−	−	x
cargarse	−	−	−	−	−	−	−	−	−	−	−	x	−	−	−	−	x	−	−
cascarse	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x
crujirse	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x
defecarse	x	−	x	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
derrumbarse	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x	−	−	−
desarmarse	x	−	x	x	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
desbordarse	x	−	−	−	−	x	−	−	−	−	−	−	−	−	−	−	−	−	−
descacharrarse	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x
descarallarse	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x
descomponerse	x	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
descoserse	−	−	−	−	−	−	−	−	−	−	−	x	−	−	−	−	x	−	−
descostillarse	x	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x
descoyuntarse	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x
desencajarse	x	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	
desfallecerse	x	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x
desgüevarse	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x
desintegrarse	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x	−	−	−
desmayarse	−	−	−	−	x	−	−	−	−	−	−	−	−	−	−	−	−	−	x
desmendrellarse	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	x

desmontarse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x
despanzarse	x	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
despanzurrarse	x	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x
desparramarse	x	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
despatarrarse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	–	x
despelotarse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x
despilrrofonarse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x
despollarse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x
destesticularse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x
destornillarse	x	–	x	–	x	–	–	–	–	–	–	x	–	–	–	x	x	–	x
desentornillarse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	–	–	–
doblarse	x	–	x	x	–	x	x	–	–	–	–	x	–	–	–	x	–	–	x
encalarse	–	–	–	–	–	x	–	–	–	–	–	–	–	–	–	–	–	–	–
escacharrarse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	x	–	–	–
escojonarse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x
escoñarse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x
espatarrarse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x
joderse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	–	–	x
matarse	x	x	x	x	–	x	–	x	x	x	–	x	–	–	–	–	–	–	x
miarse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	–	–
miccionar	–	–	–	–	–	–	–	–	–	–	–	x	–	–	–	x	–	–	–
orinarse	x	–	x	–	–	–	–	–	–	–	–	x	–	–	–	x	–	–	–
pillarse	x	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
pavonearse	–	–	–	–	–	–	–	–	–	–	–	x	–	–	–	–	–	–	–
petarse	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x
recagarse	x	–	x	x	–	–	–	–	–	–	–	–	–	–	–	x	x	–	–
recontracagarse	x	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	–	–	–
romperse	–	–	–	–	–	–	–	–	–	–	–	x	–	–	–	–	–	–	x
sacudirse	x	–	–	–	x	–	–	–	–	–	–	x	–	–	–	–	–	–	x
tentarse	x	–	x	x	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
torcerse	x	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	–	–	x
tumbarse	x	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x
zurrarse	–	–	–	–	–	–	–	–	–	–	–	x	–	–	–	–	–	–	–

Table 8. *Distribution of verbal constituents per Spanish national variety (Set B).*

Similarly to the Set A idioms, Set B idiom conglomerates can also be indicative of specific national varieties. While *cagarse de risa* can be found in ten national varieties, a combination with *caerse de risa* indicates a diatopically restricted conglomerate (Colombia, Mexico, Spain). This binary conglomerate [+cagarse/+caerse] characterises Mexican Spanish when combined with *cargarse de risa*, but only Peninsular Spanish when combined with *crujirse de risa*. In the same vein, the binary conglomerate *cagarse de risa* plus *ahogarse de risa* singles out six varieties (Argentina, Chile, Colombia, Cuba, Mexico, Peru), but in combination with *caerse de risa*, the idiom conglomerate is typical of Colombian and Mexican Spanish. Binary conglomerates also comprise negative combinations. For example, *cagarse de risa* in the absence of *caerse de risa* [+cagarse/-caerse] enables identification of two transnational varieties (Peru and Cuba); however, [+cagarse/-caerse/+atragantarse] *de risa* is diatopically restricted to Peru, whereas [+cagarse/-caerse/+desbordarse] *de risa* is marked as Cuban Spanish. In addition, idiomatic identification of Spanish national variants could be further refined by taking frequency into account.

From a metaphorical perspective, verb components of Set B idioms tend to be of a scatological nature related to body functions (cf. *mearse de risa*): *cagarse, defecarse, garcarse, miarse, miccionar, orinarse, recagarse, recontracagarse*; they make reference to the loss of sexual organs, usually male (cf. *descojonarse de risa*): *desgüevarse, despollarse, destesticularse, escojonarse, escoñarse*; or envisage laughing loud as a 'physically unmanageable' emotion leading to body upsetting (cf. *desternillarse/mondarse/retorcerse/revolcarse/tirarse de risa*), breakage (cf. *partirse/troncharse de risa*) and potential death (cf. *morirse de risa*). The latter is also one of the most productive types for Set B idioms, which elaborate figuratively on a wide range of actions and effects that indicate the inability to hold one's laughter: death (*matarse*), self-beating (*cascarse, sacudirse, zurrarse*), overwhelmingness (*atracarse, desbordarse, atacarse, encalarse*[36], *cargarse, petarse*), suffocation and blockage (*ahogarse, asfixiarse, atascarse, atorarse, atragantarse, pillarse*), disjointment (*desarmarse, descacharrarse descarallarse, descomponerse, descoserse, descostillarse, descoyuntarse, desencajarse desintegrarse, desmontarse, despanzarse, despanzurrarse, desparramarse, despatarrarse, despelotarse, despilrrofonarse, destornillarse, desentornillarse, espatarrarse*), disruption and weakness (*arrastrarse, botarse, caerse, crujirse, derrumbarse, desfallecerse desmayarse, doblarse, escacharrarse, joderse, romperse, torcerse, tumbarse*), as well as a handful of verbs not classifiable within those subcategories (*tentarse, carcajearse*[37] and *pavonearse de risa*).

5. Conclusions

Whereas standard, neutral Spanish appears as a convenient theoretical construct (though culturally and ideologically loaded as well), national varieties within the *Comunidad Panhispánica* are a powerful means of indexing social classes and building identities. Fictional discourses (literature, cinema, etc.) usually resort to the symbolic and linguo-cultural issues associated to a given national variety for character profiling and the creation of socio-cultural identitities.

It should be borne in mind that idiomatic diatopy can be signalled in many ways: through regionalisms, semantically incongruent constructions, distributional preferences

---

[36] Accoding to the DD (Deive 2002), one of the senses of the verb *encalarse* is related to spirit possession: “Introducirse un espíritu o muerto en el cuerpo de una persona” (‘(of a spirit) to enter a person’s body’ [our translation]). In such circumstances, the possessed soul is obviously overwhelmed and out of control.

[37] While the verb *carcajearse* (‘laugh’) is frequent in general Spanish, the idiom *carcajearse de risa* can only be found in the Mexican subcorpus. Cf. *reirse a carcajadas* (‘roar with laughter’) also in general Spanish.

and other frequency issues. For this reason, better resources are in great demand. Translators are faced with enormous difficulty when it comes to marking/recreating variation. Information about idiomatic diatopy is almost absent from bilingual dictionaries. Similarly, monolingual idiom dictionaries fail to provide complete information regarding distribution and frequency of idioms in national Spanish variants. The DLVEP wrongly classes Set A LOL idioms as restricted to Peninsular Spanish. In addition, one of the idioms listed (*reírse las muelas*) is not found in the subcorpora. In the same vein, no diatopic information is provided for the three LOL idioms with reflexive verbs listed under *risa* in REDES (Bosque, 2004): *morirse*, *partirse* and *troncharse de risa*; and the seven idioms in PRÁCTICO (Bosque, 2006): *morirse*, *partirse*, *troncharse*, *retorcerse, caerse*, *desternillarse* and *mondarse de risa*. More corpus-based analysis of transnational Spanish and national variants are also urgently needed.

This corpus-based case study on the rendering of the English idiom *to laugh one's head off* into Spanish has revealed a strong tendency towards 'normalising' and 'standardising' idiomaticity and diatopic variety in film subtitling. Diatopy and idiomaticity seem to undergo a double filtering: first through the 'neutral Spanish sieve' and, then, through the 'universals sieve'.

The results of the analysis show that (a) idioms in subtitling gravitate towards a 'standard' with very little room for diatopic indexing (normalisation) and that (b) avoidance of diatopic variation can even be observed in the limited choice of translation equivalents, irrespectively of the degree of idiomatic richness of the target language (and national variants).

While other idiom verbal variants are of course possible or even more frequent, below are listed all idioms that convey ‘laughing out loud’ (Sets A-B), sorted in alphabetical order with indication of national TLD.

SET A: ***descojonarse*** (.ar, .cl, .co, .mx, .pe, .uy, .ve, .es), ***desternillarse*** (.ar, .bo, .cl, .co, .mx, .ni, .pa, .py, .es, .uy, .ve), ***mearse*** (.ar, .cl, .co, .mx, .pe, .es, .uy,), ***mondarse*** (.ar, .mx, .es), ***morirse*** (.ar , .bo , .cl , .co , .cr , .cu , .do , .ec , .sv , .gt , .hn , .mx , .ni, .pa , .py , .pe , .es , .uy , .ve), ***partirse*** (.ar , .bo , .cl , .co , .cr , .cu , .do , .ec , .sv , .gt, .mx , .ni , .pa , .py , .pe , .es , .uy , .ve), ***retorcerse*** (.ar , .cl , .co , .mx , .ni , .py, .pe, .es), ***revolcarse*** (.ar , .cl , .co, .mx , .ni , .pe , .es , .uy), ***tirarse*** (.ar, .mx, .pe, .es), ***troncharse*** (.ar, .mx, .pe, .es).

SET B: ***ahogarse*** (.ar, .cl, .co, .cu, .mx, .pe), ***arrastrarse*** (.do, .sv, .pe), ***asfixiarse*** (.pe), ***atacarse*** (.mx), ***atascarse*** (.mx, .ni), ***atorarse*** (.pe), ***atracarse*** (.pe), ***atragantarse*** (.pe, .es), ***botarse*** (.mx, .ve), ***caerse*** (.co, .do, .mx, .pa, .es), ***cagarse*** (.ar , .bo , .cl , .co , .cr , .cu , .ec , .sv , .gt , .mx, .ni, .py , .pe , .es , .uy), ***carcajearse*** (.mx, .pe, .es), ***cargarse*** (.mx), ***cascarse*** (.es), ***crujirse*** (.es), ***defecarse*** (.ar, .cl), ***derrumbarse*** (.pe), ***desarmarse*** (.ar, .cl, .co), ***desbordarse*** (.ar, .cu), ***descacharrarse*** (.es), ***descarallarse*** (.es), ***descomponerse*** (.ar), ***descoserse*** (.mx, .pe), ***descostillarse*** (.ar, .es), ***descoyuntarse*** (.es), ***desencajarse*** (.ar), ***desfallecerse*** (.ar, .es), ***desgüevarse*** (.es), ***desintegrarse*** (.pe), ***desmayarse*** (.cr, .es), ***desmontarse*** (.es), ***despanzarse*** (.ar), ***despanzurrarse*** (.ar, .es), ***desparramarse*** (.ar), ***despatarrarse*** (.es, .uy), ***despelotarse*** (.es), ***despilrrofonarse*** (.es), ***despollarse*** (.es), , ***destesticularse*** (.es), ***destornillarse*** (.ar, .cl, .cr, .mx, .pe, .es, uy), ***desentornillarse*** (.pe), ***doblarse*** (.ar, .cl, .co, .cu, .do, .mx, .pe, .es), ***encalarse*** (.cu), ***escacharrarse*** (.py, .pe, .es), ***escojonarse*** (.es), ***escoñarse*** (.es), ***espatarrarse*** (.es), ***garcarse*** (.ar, .uy), ***joderse*** (.pe, .es), ***matarse*** (.ar, .bo, .cl, .co, .cu, .ec, .sv, .gt, .mx, .es), ***miarse*** (.uy), ***miccionar*** (.mx, .pe), ***orinarse*** (.mx, .pe), ***pillarse*** (.ar), ***pavonearse*** (.mx), ***petarse*** (.es), ***recagarse*** (.ar, .bo, .cl, .py, .uy), ***recontracagarse*** (.ar), ***romperse*** (.mx, .es), ***sacudirse*** (.ar, .cr, .mx, .es), ***tentarse*** (.ar, .cl, .co), ***torcerse*** (.ar, .pe, .es), ***tumbarse*** (.ar, .es), ***zurrarse*** (.mx).

Out of 74 synonymous idioms with the pattern [V_{reflex} + DE RISA] that are available for Spanish according to the comparable corpora analysed, only four appear in the parallel corpus of film subtitles: 1. *morirse de risa*, 2. *partirse de risa*, 3. *desternillarse de risa* (Set A) and 4. *cagarse de risa* (Set B). The first three Set A idioms occupy identical, top positions in general Spanish (and closer American Spanish): 1. *morirse*, 2. *partirse*, 3. *desternillarse de risa*, whereas the Peninsular variety features 1. *partirse*, 2. *morirse* and 5. *desternillarse de risa* (Table 4). As to distribuition, *morirse de risa* can be found in all 19 national varieties, *partirse de risa* appears in 16 varieties, *desternillarse de risa*, in 13 national varieties, and *cagarse de risa* (Set B) is documented in 15 national varieties (cf. Table 7). Those four LOL idioms are, then, widely and frequently used in most Spanish national varieties. Therefore, they could class as ‘standard’ transnational Spanish.

No diatopically restricted equivalent idioms seem to have been favoured in the bilingual corpus of film subtitles in English and Spanish. Similarly, there is no trace of the figurative richness of Set B idioms within the textual equivalents retrieved from the parallel corpus.

However, this fact does not necessarily mean that idiomatic diatopy is completely ‘removed’ in translation. One interesting finding of our study is the existence of divergent configurations (vectors of features or conglomerates) based on idioms distribution, frequency and co-selection which could be indicative of specific national varieties. Studying idiom conglomerates to detect diatopy is a promising avenue of research which could definitely benefit from giga-token corpora and NLP tools. This type of data could successfully enrich other corpus-based translation proposals, like the Model of Dialect Reconstruction (Sánchez Galvis 2012). In addition, this type of analysis could be also successfully applied in teaching translation (of subtitles) or developing CAT tools for subtitling.

A word of caution is needed, though. Idioms, as opposed to collocations, are low-frequency items. For this reason, extremely large corpora are required. Web-crawled giga-token corpora are presently providing invaluable data, but they are not free from technical limitations. Problems with corpus preparation and processing (document selection, cleaning-up, deduplication), parsing and annotation errors, as well as punctuation errors, substandard spellings and typos, could seriously compromise results.

6. Acknowledgements

The research presented in this paper has been partially carried out in the framework of the research projects EXPERT (317471-FP7-PEOPLE-2012-ITN), INTELITERM (FFI2012–38881) and TERMITUR (HUM2754).

7. References

Agost Cano, Rosa. 1998. “La importancia de la variació lingüística en la traducció.” *Quaderns. Revista de traducción* 2: 83–95.

Bolaños Cuéllar, Sergio. 2004. “Sobre los límites de la traducibilidad: la variación dialectal textual.” *Íkala, revista de lenguaje y cultura* 9 (15): 315–347.

Benko, Vladimír. 2014. “Aranea: yet another family of (comparable) web corpora.” In *Text, speech and dialogue. 17th international conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings* (LNCS 8655), ed. by Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala 257–264. Heidelberg, New York, Dordrecht and London: Springer.

Benko, Vladimír. 2015a. *Araneum Hispanicum Maius*, version 15.04. Ústav Českého národního korpusu FF UK, Prague 2015. Dostupný z. https://www.korpus.cz (ARANEUM_MA).

Benko, Vladimír. 2015b. *Araneum Hispanicum Minus*, version 15.04. Ústav Českého národního korpusu FF UK, Prague 2015. Dostupný z. https://www.korpus.cz (ARANEUM).

Bosque, Ignacio (dir). 2004. *Redes. Diccionario combinatorio del español contemporáneo.* Madrid: Ediciones SM (REDES).

Bosque, Ignacio (dir). 2006. *Diccionario combinatorio práctico del español contemporáneo*. Madrid: Ediciones SM (PRÁCTICO).

Bravo, José María. 2006. "Film Translation Research in Spain." In *Lexicography, Terminology, and Translation: Text-based Studies in Honour of Ingrid Meyer*, ed. by Ingrid Meyer, and Lynne Bowker. 227–237. Ottawa: University of Ottawa Press.

Brutt-Griffler, Janina. 2002. *World English: A Study of Its Development*. Clevedon, UK: Multilingual Matters.

Corpas Pastor, Gloria. 1995. "Discoursal Functions of Proverbs. A Corpus-based Study." *Estudios Ingleses de la Universidad Complutens*e 3: 101–110.

Corpas Pastor, Gloria. 1999. "¿Cómo traducir las variedades dialectales?" In *Actas del VI Simposio Internacional de Comunicación Social. Santiago de Cuba, 25-28 de enero de 1999*. 2 volumes, ed. by Leonel Ruiz Miyares. 1233–1239. Santiago de Compostela: Centro de Lingüística Aplicada/ Oriente: Consiglio Nazionale delle Ricerche. 1233–1239.

Corpas Pastor, Gloria. 2008. *Traducir con corpus: los retos de un nuevo paradigma*. Frankfurt: Peter Lang.

Corpas Pastor, Gloria. 2013. "Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas," In *Fraseopragmática*, ed. by Inés Olza, and Elvira Manero, 335–373. (Romanistik). Berlin: Frank & Timme.

Corpas Pastor, Gloria. 2015. "Translating English Verbal Collocations into Spanish: on Distribution and other Relevant Differences related to Diatopic Variation." *Lingvisticæ Investigationes* 38 (2): 229–262. Special

Issue “Spanish Phraseology. Varieties and variations”, ed. by Pedro Mogorrón Huerta.

Corpas Pastor, Gloria, Johanna Monti, Ruslan Mitkov, and Violeta Seretan (eds). 2013. *Workshop Proceedings for: Multi-word units in Machine Translation and Translation Technologies*. Allschwil (Switzerland): European Association for Machine Translation (EAMT). http://www.mtsummit2013.info/files/proceedings/WkSh4_proceedings.pdf>.

Davies, Mark. 2002-. *Corpus del Español: 100 million words, 1200s-1900s*. (BYU-Davies). http://www.corpusdelespanol.org.

Deive, Carlos E. 2002. *Diccionario de Dominicanismos*. 2nd ed. Santo Domingo: Ediciones Librería La Trinitaria / Editora Manatí (DD).

Dennisov, Stephanie (ed). 2013. *Contemporary Hispanic Cinema: Interrogating the Transnational in Spanish and Latin American Film*. New York, NY: Tamesis Books.

Díaz-Campos, Manuel (ed). 2015. *The Handbook of Hispanic Sociolinguistics*. Chichester, West Sussex: John Wiley & Sons (paperback edition).

Díaz Cintas, Jorge. 2009. “Introduction – Audiovisual Translation: An Overview of its Potential.” In *New Trends in Audiovisual Translation*, ed. by Jorge Díaz Cintas, 1–20. Bristol: Multilingual.

Feldman, Anna, and Jing Peng. 2013. “Automatic detection of idiomatic clauses.” In *CICLing'13 Proceedings of the 14th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*. 435–446. Berlin: Springer.

Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

Heid, Ulrich. 2008. “Computational Phraseology: an Overview.” In *Phraseology. An Interdisciplinary Perspective,* ed. by Silvianne Granger and Fanny Meunier, 337–360. Amsterdam and Philadelphia: John Benjamins.

Hualde, José Ignacio, Antxon Olarrea and Erin O’Rourke (eds). 2012. *Handbook of Hispanic Linguistics*. Chichester, West Sussex: John Wiley & Sons.

Ilisei, Iustina, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. “Identification of Translationese: A Supervised Learning Approach.” In *Computational Linguistics and Intelligent Text Processing, 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010. Proceedings.* (Lecture Notes in Computer Science 6008). 503–511. Heidelberg: Springer.

Kachru, Braj B. 1985. “Standards, codification and sociolinguistic realism: the English language in the outer circle.” In *English in the world: Teaching and learning the language and literatures*, ed. by Randolf Quirk and H.G. Widdowson, 11–30. Cambridge: Cambridge University Press.

Kachru, Braj B. (ed). 1992. *The Other Tongue: English across Cultures*. 2nd edition. Urbana and Chicago: University of Illinois Press.

Kachru, Braj B. 1997. “World Englishes and English-using communities.” *Annual Review of Applied Linguistics* 17: 66–87.

Kachru, Braj. B., Yamuna Kachru, and Cecil L. Nelson (eds). 2006. *The Handbook of World Englishes*. Oxford: Blackwell.

Kenny, Dorothy. 2014. *Lexis and Creativity in Translation: A Corpus-Based Study*. Oxon and New York: Routledge.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. “The Sketch Engine: ten years on.” *Lexicography: Journal of ASIALEX* 1 (1): 7–36.

Kirkpatrick, Andy. 2007. *World Englishes: Implications for International Communication and English Language Teaching* (Cambridge Language Teaching Library). Cambridge: Cambridge University Press.

Laviosa, Sara. 2002. *Corpus-based Translation Studies. Theory, Findings, Applications*. Amsterdam and New York: Rodopi.

Lawick, Heike van. 2007. “Phraseologie und Übersetzung unter Anwendung von Parallelkorpora. ” In *Translation Studies. Doubts and Directions,* ed. by Miriam Schlesinger and Radegundis Stolze, 281–296. Amsterdam and Philadelphia: John Benjamins.

Leppihalme, Ritva. 2000. “Päätalo Idioms and Catchphrases in Translation.” In *Erikoiskielet ja käännösteoria*. *VAKKI-symposiumi XX*. Vaasa: University of Vaasa. 224–234.

Lipski, John. M. 2012. “Geographical and Social Varieties of Spanish: An Overview.” In *Handbook of Hispanic Linguistics,* ed. by José Ignacio Hualde, Antxon Olarrea, and Erin O’Rourke. 1–26. Chichester, West Sussex: John Wiley & Sons.

Lison, Pierre and Jörg Tiedemann. 2016. “OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles.” In *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2016), ed. by Nicoletta Calzolari, Khalid Choukri, Thierry

Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk and Stelios Piperidis, 923–929. Paris: European Language Resources Association.

Marco Borillo, Josep. 2009. “Normalisation and the translation of phraseology in the COVALT corpus.” *Meta* 54 (4): 842–856.

Marco Borillo, Josep. 2011. “Some Insights into the Factors Underlying the Translation of Phraseology in the COVALT corpus.” In *Beyond Borders – Translations Moving Languages, Literatures and Cultures*, ed. by Pekka Kujamäki, Leena Kolehmainen, Esa Penttilä, and Hannu Kemppanen. 197–214. Berlin: Frank & Timme.

Mar-Molinero, Clare and Paffey, Darren. 2015. “Linguistic Imperialism: Who Owns Global Spanish?” In *The Handbook of Hispanic Sociolinguistics,* ed. by Manuel Díaz-Campos. 747–764. Chichester, West Sussex: John Wiley & Sons.

Mauranen, Anna and Pekka Kujamäki (eds). 2004. *Translation Universals: Do They Exist?* Amsterdam: John Benjamins.

Mayoral Asencio, Roberto. 1999. *La traducción de la variación lingüística*. (*Uertere. Monográficos de la revista Hermeneus* 1). Soria: Diputación Provincial de Soria.

McArthur, Tom. 2001. World English and world Englishes: Trends, tensions, varieties, and standards. *Language Teaching* 34: 1–20.

Niño-Murcia, Mercedes. 2015. “Variation and Identity in the Americas.” In *The Handbook of Hispanic Sociolinguistics,* ed. by Manuel Díaz-Campos. 728–746. Chichester, West Sussex: John Wiley & Sons.

Oakes, Leigh 2001. *Language and National Identity*. Amsterdam and Philadelphia: John Benjamins.

Paffey, Darren. 2012. *Language Ideologies and the Globalisation of 'Standard' Spanish.* London and New York: Bloomsbury.

Penadés Martínez, Inmaculada. 2002. *Diccionario de locuciones verbales para la enseñanza del español*. Madrid: Arco Libros. [DLVEP]

Pomikálek, Jan. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. PhD dissertation, University of Brno. http://is.muni.cz/th/45523/fi_d/phdthesis.pdf.

Pym, Anthony. 2008. "On Toury's laws of how translators translate." In *Beyond Descriptive Translation Studies: Investigations in Homage to Gideon Toury*, ed. by Anthony Pym, Miriam Shlesinger and Daniel Simeoni. 311–328. Amsterdam and Philadelphia: John Benjamins.

Ramos Pinto, Sara. 2009. "How important is the way you say it? A discussion on the translation of linguistic varieties." *Target* 21 (2): 289–307.

Real Academia Española (n.d.). Banco de datos (CORPES XXI) [online]. Corpus del español del siglo XXI. http://www.rae.es

Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. "Multiword Expressions: a Pain in the Neck for NLP." In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, CICLing'02*, 1–15. London: Springer.

Sánchez Galvis, Jairo A. 2012. "Traducción y variedad lingüística: hacia un modelo de reconstrucción textual." *Revista electrónica de lingüística aplicada* 11: 125–136.

Schäfer, Rolan, and Felix Bildhauer. 2012. "Building Large Corpora from the Web Using a New Efficient Tool Chain." In *Proceedings of LREC'12*, 486–493.

Schäfer, Rolan, and Felix Bildhauer. 2013. *Web corpus construction* (Synthesis Lectures on Human Language Technologies). San Francisco: Morgan & Claypool.

Subirats, Carlos, and Marc Ortega. 2012. *Corpus del Español Actual* (CEA) http://spanishfn.org/tools/cea/english.

Seargeant, Philip. 2012. *Exploring World Englishes: Language in a Global Context*. New York and Oxon: Routledge.

Suchomel, Vít, and Jan Pomikálek. 2012. "Efficient web crawling for large text corpora." In *Proceedings of the seventh web as corpus workshop (WAC7),* ed. by Adam Kilgarriff and Serge Sharoff, 39–43. Lyon.

Sumillera, Rocío. 2008. "Postcolonialism and Translation: the Translation of *Wide Sargasso Sea* into Spanish." *New Voices in Translation Studies* 4: 26–41.

Tiedemann, Jörg. 2009. "News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces." In *Recent Advances in Natural Language Processing V. Selected Papers from RANLP 2007*, ed. by Nicolai Nicolov, Kalina Bontcheva, Galia Angelova and Ruslan Mitkov, 237–248. Amsterdam and Philadelphia: John Benjamins.

Valdés, Guadalupe, and Michelle Geoffrion-Vinci. 2012. "Heritage Language Students: The Case of Spanish." In *Handbook of Hispanic Linguistics*, ed. by José Ignacio Hualde, Antxon Olarrea, and Erin O'Rourke. 598–622. Chichester, West Sussex: John Wiley & Sons.

del Valle, José (ed). 2007. *La lengua, ¿patria común? Ideas e ideologías del español*. Madrid and Frankfurt: Iberoamericana and Vervuert.

del Valle, José. 2011. "Transnational languages: beyond nation and empire? An introduction." *Sociolinguistic Studies* 5 (3): 387–397.

Wright, Sue. 2004. *Language policy and language planning: From nationalism to globalization.* New York: Palgrave Macmillan.