# Can comparable corpora be compared?

**Belén López Arroyo**
ACTRES-Universidad de Valladolid (Spain)
belenl@lia.uva.es

## Abstract

While there is consensus on the definition of a comparable corpus, there is little or no agreement on what makes a corpus comparable or how to assess comparability. A comparable corpus consists of two or more collections of texts (subcorpora) in different languages or different language varieties, which are similar in some way. But in what way? According to McEnery and Xiao (2007: 20), proportion, genre, domain, and time constitute the main criteria when compiling a comparable corpus and must match in the different languages for the corpus to be considered comparable. However, in previous studies (López-Arroyo & Roberts, 2017), it has been shown by the analysis of two specialized comparable corpora that these criteria work well for certain fields, but not all. In the present study, we examine comparability from the point of view of the purpose for which a comparable corpus is to be used. In order to do that we have compiled a comparable corpus of 150 tasting notes in English and Spanish written by two experts in the field in Spain and in usa and published in the same decades; according to McEnery and Xiao (2007) our corpora meet all the criteria to be comparable. However, our methodology focused on the analysis of aspects such as content, format and style of the genre under study for the comparability of corpora will prove that proportion, genre, domain, time and size are not valid enough for comparing comparable corpora.

**Keywords:** comparable corpus, comparability, purpose, tasting notes.

## Resumen

***Sobre la comparación de corpus comparables***

Podemos afirmar que, hoy en día, no existe un acuerdo unánime sobre los criterios para compilar un corpus comparable o sobre cómo evaluar la comparabilidad de un corpus. Un corpus comparable es una colección de textos en diferentes lenguas o variaciones que son similares en ciertos aspectos. Pero, ¿en cuáles? Según McEnery y Wilson (2007: 20), la proporción en las muestras, el género, campo y tiempo deben ser los criterios principales a la hora de

compilar un corpus comparable y deben ser los mismos en las diferentes lenguas. Sin embargo, estudios previos (López-Arroyo & Roberts, 2017) demuestran que estos criterios pueden no ser válidos en todos los campos. En el presente estudio, analizamos la comparabilidad desde el punto de vista del propósito del corpus. Para ello, hemos compilado un corpus comparable de 150 fichas de cata en inglés y 150 en español escritas por dos autoridades del campo y publicadas en las mismas décadas; según McEnery y Xiao (2007) nuestros subcorpus reúnen todos los requisitos para ser comparables. Sin embargo, nuestra metodología, centrada en el análisis de otros factores tales como El formato, el contenido y el estilo, demostrará que únicamente la proporción, el género, el campo, el tiempo y el tamaño no son siempre suficientes a la hora de comparar corpus.

**Palabras clave:** corpus comparable, criterios de comparación, propósito, fichas de cata.

## 1. Introduction

"A comparable corpus is one which selects similar texts in more than one language or variety," according to The EAGLES - Expert Advisory Group on Language Engineering Standards Guidelines (1996). The EAGLES report considers that a comparable corpus allows linguists to compare different languages or varieties in similar circumstances of communication, but avoids the inevitable distortion introduced by the translations found in a parallel corpus. However, for a comparable corpus to be an effective tool, the texts must be "similar." And, as the EAGLES group stated in 1996, "There is as yet no agreement on the nature of the similarity" (1996: 12).

This issue of similarity was addressed to some extent by McEnery and Xiao (2007: 20), who stated that multilingual comparable corpora are to include "the same proportions of the texts of the same genres in the same domains in a range of different languages in the same period". In other words, according to these authors, proportion, genre, domain, and time constitute the main criteria when compiling a comparable corpus and must match in the different languages for the corpus to be considered comparable. The problem is that these external criteria do not always guarantee that the different language subcorpora in a comparable corpus match, because the same genre may be used to address a different audience or may show different features in its construction in different languages. In fact, the same genre, when compared interlinguistically, may show differences in content or in style, making it difficult to draw valid contrastive conclusions (López-Arroyo & Roberts, 2016: 399).

In a previous study (López-Arroyo & Roberts, 2017), two specific purpose comparable corpora built using two genres (abstracts and wine tasting notes) were examined, to see whether the criterion of genre, supplemented by the criteria of proportion, domain, and time, is sufficient to build comparable corpora, as McEnery and Xiao (2007) have suggested. The analysis of the abstracts and wine tasting notes corpora showed that the criteria for compilation of a comparable corpus proposed by McEnery and Xiao (genre, proportion, domain and time) work well for certain fields, but not all, and that the degree of comparability varies from one corpus to another.

In the present study, we will examine the comparability of a comparable corpus from the point of view of the purpose for which it is being used. In other words, we will attempt to answer the following question: is a given comparable corpus comparable enough to meet the researcher's specific needs?

We will begin by examining the concept of comparability and by reviewing different methods of assessing the comparability of comparable corpora. Then, we will present our own corpora: a comparable corpus of 150 tasting notes in English and Spanish written by two experts in the field in Spain and in USA and published in the same decade; both corpora have the same number of samples; according to McEnery and Xiao (2007) our corpus meets all the criteria to be comparable. However, our methodology states that other criteria should be taken into to the consideration for the comparability of comparable corpora; the analysis of aspects such as content, format and style of the genre under study will prove that genre, domain, time and size are not valid enough for comparing comparable corpora. Our second research question is, are corpora always comparable?

## 2. Concept of comparability

We can say a corpus is a comparable corpus if its components or subcorpora are collected using the same sampling frame and similar balance and representativeness (McEnery, 2003: 450); for example, the same proportions of the texts of the same genres in the same domains in a range of different languages (or language varieties) in the same sampling period. However, the subcorpora of a comparable corpus are not translations of each other. Rather their comparability lies in their same sampling frame, which allows for the collection of similar texts.

The concept of comparability is thus related to that of similarity, which is a concept characterized by vagueness. Moreover, the concept of similarity in terms of corpora is a complex one. One has to pursue this notion of similarity across languages, cultures, text varieties or genres, and circumstances of communication. And, since we are talking about corpora – as opposed to just texts – we must consider similarity in relation to both form and content. The form of corpora will be of interest to those who actually construct them. In other words, one needs to consider the size of the corpus, in terms of number of texts and words, and the nature of the individual texts in terms of words, sentences, and paragraphs. The content of corpora is of paramount importance in the construction of comparable corpora, and guidelines on what to look for should be drawn up for this purpose. One needs to start by looking at the similarity of the texts in relation to their structure, their style and their function. Is the structure important? For example, are they formal, carefully constructed texts – as with legal texts, or are they informal, loosely organized discourse – as with transcriptions of conversation. Then one needs to look at their function in the source culture. Do the texts have a specific social or cultural importance that needs to be taken into consideration? All these considerations complicate the concept of similarity and comparability.

As Sharoff points out (2013: 113), "the notion of comparable corpora rests on our ability to assess the difference between corpora which are claimed to be comparable, but this activity is still art rather than proper science." Despite the quantitative measures employed to detect similarities and differences between subcorpora, there is no consensus on how to assess the comparability of a comparable corpus.

This led Maia (2003: 34) to conclude that "comparability is in the eye of the beholder". This is not a satisfactory state of affairs, for we do not want to base important linguistic conclusions on sources that may be subjectively selected. We could avoid subjectivity if we could assess how comparable, or similar, two subcorpora are.

## 3. Comparable corpora and their purpose

Li (2012: 35) suggests that "the way we define comparability should depend on the target applications in consideration," since different applications might prefer comparable corpora of different genres. For example, Ni et al.

(2009) built multilingual corpora from Wikipedia for use in cross-lingual text classification (using texts labeled in one language to help classify texts in another language) and cross-lingual document reference (matching a given document in one language with related documents in another language). To serve these applications, they used a group of "universal" topics to model documents from different languages. However, such topic modeling is not required for comparable corpora to be used for bilingual lexicon extraction or for contrastive analysis of grammatical features.

Our study is based on the same premise, that the definition of comparability and the degree of comparability required in a comparable corpus depends on the purpose for which it is established, the use to which it is to be put; apart from that, we also want to test whether comparable corpora are always comparable.

## 4. Compilation of the English/Spanish comparable corpus: wine tasting notes

The basis of our study is an English/Spanish comparable corpus of wine tasting notes. Wine tasting notes, a genre used in oenology, are standardized texts, used either during professional wine tasting or when new wines are released, to record the different organoleptic features or components of a wine. Wine tasting notes can be considered as a subsidiary genre of Wine Tasting Technical Sheets since they often form part of the latter and have the same function, one that is relevant for the professional discourse community. Wine tasting notes are short texts typically organized in three different sections that correspond to the three steps in any wine tasting procedure: "the assessment of wine's colour, its smell (metonymically referred to as its *nose* / *nariz* in English and Spanish), and its mouth-feel (a stage that involves smell, taste, and touch, is metonymically referred to as the wine's *palate* / *boca*, and may be 'de-composed' into several stages" (Caballero, 2017: 69).

Peynaud (1987), distinguishes three discourse communities dealing with tasting notes according to the writer: the professional taster, the amateur presenting a wine to guests at a dinner tasting, and a wine journalist writing for the readers of a wine magazine. He goes on to discuss several ways of talking about the taste of wine depending on circumstances, training and the taster's state of mind.

It seems obvious that such a variety of types of writers will produce different types of tasting notes that could not be considered comparable because of their differences in style, the format and the way to render knowledge to readers.

### 4.1. Corpus compilation

Considering the facts stated above, our starting point was to focus on tasting notes written by one type of discourse community in English and Spanish so as to be able to adjust comparability. We restricted our corpus to tasting notes written by experts. The expert seeks clarity and precision above all in his expression. His style is strict and economical but his comments are reasoned; his conciseness is not due to a lack of imagination but to a choice of the most precise words, and in his reports he only uses terms with an accepted and agreed meaning. In spite of his skills, his language should be simple and intelligible to all. Where technical terms are concerned he will refrain from defining smells by analogy with little known chemical substances (López-Arroyo & Roberts, 2017: 375).

We compiled a corpus of 300 tasting notes produced by the best known and most influential wine critics writing in the two languages under study here: for English, Robert Parker, an American wine critic with an international reputation, who developed a 100-point wine-scoring scale; for Spanish, José Peñín, the most successful wine writer and journalist in the Spanish speaking world today. Both these wine writers warrant a corpus of their own not only because of their importance in the wine tasting world, but also because of the distinctive style of their tasting notes.

It could be argued that by compiling a corpus of wine tasting notes written by only two authors, the results could be more concerned with the analysis of their styles rather than the representativity of the genre; however, Parker and Peñín represent a group of wine critics in English and Spanish respectively who publish thousands of wine tasting notes a year and address a vast audience. In the present paper, we take the samples written by Parker and Peñín as representatives of the wine critics in journals such as *The Advocate* and Guía Peñín.

# 5. Methodology and results

As we can see, our corpus meets McEnery and Xiao's (2007) criteria for comparing corpora. Same genre and domain (wine tasting notes), same period of time (2005-2015) and same participants (wine critics writing their own publication). However, our starting point was that those criteria do not seem to be always enough for a corpus to be comparable. That is why our methodology takes one step further and analyzes the samples in terms of format, style and content to test the degree of comparability between the two corpus under study.

## 5.1. Differences in content

Despite the apparent consensus that all tasting notes cover at the very least the colour, nose, and palate of wines (Caballero, 2007, 2009, 2017; Caballero & Suarez-Toste, 2008: 383), Wipf (2010: 15), using a concrete example, rightly points out that this is not always the case: "certain parts are left out, whereas others are added".

However, it is rare to find a tasting note that covers all the above moves and steps. A previous study on the content of tasting notes (López-Arroyo & Roberts, 2016) written by experts, journalists publishing in wine journals and wineries revealed that, in English, if a key move is dropped, it is invariably Appearance (66% of the samples did not include the section Appearance).

So, the next step is to see a) whether that is the case in the wine critics corpus and b) if it is left out more in one language than in the other.

| EN | Total no. of tasting notes | No. of tasting notes without Appearance |
|---|---|---|
| Robert Parker subcorpus | 150 | 42 (66%) |

Table 1. Number of tasting notes without the Appearance move in the English corpus.

Table 1 above shows that only 28% of the tasting notes in our English corpus include Appearance. While one might wonder why Parker seems to buck the trend in English (see above) to gloss over the appearance of a wine, there is a simple explanation for his inclusion of colour in most of his tasting notes. In the early 1970s, he devised a system to taste and score wines. And, although he claims that this is not a scientific system, it looks at the colour, bouquet and taste of a wine, each of which is worth a certain number of

points. By awarding five points for the colour, 15 for the bouquet and 20 for the palate and texture, on top of a base score of 50 that each wine starts with, Parker built a numerical scale that not only helped demystify wine for a generation of budding oenophiles but would go on to become the most powerful rating system in the world of wine. Inclusion of colour, then, is necessary for his scoring system to work properly. However, even Parker does not always cover colour in the verbal expression of the wine tasting process, i.e. in the description of the wine in the tasting note.

Is the same true in Spanish? (See table 2 below).

| EN | Total no. of tasting notes | No. of tasting notes without Appearance |
|---|---|---|
| José Peñín subcorpus | 150 | 3 (2%) |

Table 2. Number of tasting notes without the Appearance move in the Spanish corpus.

Only 2 % of the Spanish samples do not include Appearance, which indicates that overall Appearance figures far more frequently in Spanish wine tasting notes than in the English ones. Peñín's systematic inclusion of appearance could perhaps be explained by the fact that he too, like Parker, uses a 100-point scoring scale to grade wines and therefore needs to explicitly include the visual aspect. But the similarity among his tasting notes suggests another explanation: that Peñín uses a template to prepare his notes, a template that starts with colour.

In summary then, appearance is most often omitted in Parker's tasting notes, which is not the case in Spanish. It is interesting to note that even though both Parker and Peñín include Appearance in their own method to evaluate wines, Parker does not follow it on a regular basis (66% of the samples did not include it). Apart from that, it is also interesting to see that the best known writers of wine tasting notes in English and in Spanish, Parker and Peñín, both adhere in their notes to the presentation of the three steps of the wine tasting process - analysis of appearance, aroma and taste – but that their influence on their respective colleagues in this regard is not obvious, as other wine critics writing in English and Spanish wine journals often omit the appearance move much more than they do.

### 5.2. Differences in format

In a previous study (López-Arroyo & Roberts, 2016), which was based on a

corpus of 100 English notes and 100 Spanish notes produced by wineries, wine critics and wine journalists, we had noted two different formats used to write these notes. Some notes were produced in the form of a paragraph of running text, with no clearcut divisions between the various parts of the note. Others separated the main parts of the note (Appearance, Aroma, Taste) and distinguished them clearly by using subheadings.

We also noticed that the divided format was used much more frequently in Spanish than in English. All this led us to wonder if one format was more favoured in one language than the other (see table 3 below).

| EN | Total no. of tasting notes | No. of tasting notes using running text format | No. of tasting notes using divided format |
|---|---|---|---|
| Robert Parker subcorpus | 150 | 150 (100%) | 0 |

Table 3. Formatting of wine tasting notes in the English corpus.

While, in the light of our previous study, we were not expecting a large number of English notes with divided format, now we were nonetheless surprised not to find a single one. However, the lack of divided format notes can perhaps be explained in terms of certain stylistic features which will be analyzed in the following section.

| EN | Total no. of tasting notes | No. of tasting notes using running text format | No. of tasting notes using divided format |
|---|---|---|---|
| José Peñín subcorpus | 150 | 0 | 150 (100%) |

Table 4. Formatting of wine tasting notes in the Spanish corpus.

It is surprising to see that all the samples in the Spanish corpus (150) use the divided format.

Example (1) shows an example of Parker tasting notes where no appearance and no divided format are included:

(1) Subtle and racy, with lemon rind, vanilla cream and dried pineapple. Very spicy and intense. Full-bodied, with great length and flavor. Electrified yet refined, with medium sweetness and a wonderful finish. I love the class of this, and the length. Has afterburners.

Here is an example of Peñín's tasting notes:

(2) Color cereza borde granate. Aroma intensidad media, hierbas de tocador, fruta escarchada. Boca sabroso, especiado, taninos maduros. (Color Cherry, deep red rim. Aroma Medium intensity, dressing table herbs, candied fruit. Taste Savoury, spiced, ripe tannins).

While Peñín does not start each major part on a separate line, it is clear that he has 3 divisions, which he entitles *Color*, *Aroma* and *Boca*. And although these subtitles are, strangely enough, not separated from what follows by a colon, they are clearly meant to be subtitles. The fact that all of Peñín's tasting notes follow exactly the same pattern further leads us to believe that this wine taster is using a prepared template that consists of the three subtitles and that he adds a few words after each subtitle to compose his tasting note.

In summary, the divided format is popular in the Spanish corpus, whereas the running text format is definitely the preferred option for Parker. Considering this opposite tendency, can we consider that both corpora are comparable in format?

### 5.3. Differences in style

However, we feel that the greatest differences between the two corpora of tasting notes are at the level of style. Style is seen here as the conscious or unconscious selection of a set of features from all the possibilities in a language. It is considered as any situationally distinctive use of language – a characteristic of groups as well as of individuals. (Crystal, 1987: 66).

Stylistic differences are apparent in examples three and four presented below:

(3) This was a bucket of jammy berries, vanilla and cooked sugar. With the texture of motor oil, heat like rocket fuel and flavors of a jammy berry pie, one sip was two sips too much. Perhaps as a dessert sauce, poured over vanilla ice cream, this would work. But as a wine, it was not my style. The wine was poured double blind by one of my friends thinking I would like it, if I did not know what it was. He was wrong. While I can see why some people would like this wine, (Which is why it scored 80) I am positive the wine and I will both be better off if we do not meet again.

(4) Color cereza claro. Aroma fruta fresca, intensidad media, franco. Boca fresco, frutoso, sabroso. (Color Light cherry. Aroma Fresh fruit, medium intensity, honest. Taste Fresh, fruity, savoury).

The first difference that drew our attention was that Peñín's tasting notes are verbless. He defines wines using almost a telegraphic style, just with nouns and adjectives. On the other hand, Parker tends to use a more complex discourse or style, addressing the reader, expressing openly his opinion, and using superlatives. These facts made us expand our criteria to analyse tasting notes in terms of the style authors used to see the extent in the differences between the two subcorpora.

As stated above, while the stylistic differences can, to some extent, be attributed to individual choices, our preliminary analysis of the comparable corpus led us to believe that there were characteristic features distinguishing different categories of tasting notes in the two languages under study.

In fact, we identified six distinctive features that appeared in the two corpora examined and which are listed below:

1. Superlatives
2. Figurative language
3. Personal intervention
4. Addressing the reader
5. Conversational style

Each of these will be defined, analyzed and discussed in the following sections.

### 5.3.1. Superlatives

Tasting notes — and those who write them — are often mocked, because the language can be too flowery, the descriptions overblown, the flavours impossible to believe. The first stylistic feature we therefore wish to examine, the use of superlatives, reflects this tendency towards hyperbole and exaggeration. We wish to see if this feature is found in most tasting notes or mainly in tasting notes in English.

The term "superlatives" is used here as a generic term to cover not only the grammatical sense of the superlative expressions or forms of adjectives or adverbs, but also something embodying the highest form of a thing, as well as expressions of abundant praise. It includes the following elements:

(i) Strict superlatives in the grammatical sense (highest, the most measured)

(ii) Words expressing intensity in the broadest sense of the word (quintessential, extraordinarily, pure perfection)

(iii) Different ways of designating a wine: (this beauty, this youngster).

(iv) Words expressing the concept of the wine being incomparable: (historic, classic)

(v) Words denoting the concept of plenty, of weight, of power: (massive, loads of)

Obviously, there are other elements that could fit into the category of superlatives, but we felt that the five listed above covered the category adequately.

Each of these elements was examined in each of the subcorpora, using either formal patterns (est for English superlative adjectives) or specific lexical items. The list of items searched in English and the results are found in table 5 below:

| Element examined | Items searched | Robert Parker Subcorpus |
|---|---|---|
| Strict superlatives | -est | 2 |
| | the most | 14 |
| | Total | 16 |
| Words expressing intensity | quintessential | 2 |
| | extraordinary | 15 |
| | perfect | 4 |
| | exceptionally | 3 |
| | extraordinarily | 2 |
| | perfectly | 2 |
| | (pure/near) perfection | 2 |
| | Total | 30 |
| Different ways of designating a wine | this beauty | 3 |
| | this/a modern-day legend | 2 |
| | Total | 5 |
| Words expressing the concept of the wine being incomparable | historic | 1 |
| | classic | 8 |
| | never seen before | 1 |
| | Total | 10 |
| Words denoting the concept of plenty, of weight, of power | massive | 8 |
| | loads of | 2 |
| | Total | 10 |
| | **Total of all superlatives** | **71** |

Table 5. Results of the analysis of superlatives in the English corpus.

What does this table reveal? First and foremost, it shows a clear difference in the use of superlatives from one corpus to the next: the Parker corpus is prolific in their use. This positioning of the corpus in relation to superlatives is understandable from one point of view: wine critics generally see themselves as guides, "leading readers on a quest to explore what is most beautiful, fascinating, distinctive, curious, delicious and moving in wine" (Asimov, 2014). They are generally not wine scientists who feel a certain obligation to stay neutral and objective. So, it makes sense that the Parker corpus uses superlatives to define wine. What is somewhat surprising, however, is the fact that Robert Parker, inventor of the seemingly objective 100-point wine scoring scale, is most given to using superlatives, although he has admitted that emotions do matter, contrary to the apparent objectivity of the 100-point scale: "I really think probably the only difference between a 96-, 97-, 98-, 99- and 100-point wine is really the emotion of the moment" (Tobley-Martínez, 2007: 46). There is no doubt that Parker shows his emotions through the use of superlatives in his tasting notes. He admitted as much when he said: "As a consumer advocate, you are required, expected, to express your opinion and the consumer can agree or disagree. Do I sometimes overdo it and get carried away? No doubt about it" (Tobley-Martínez, 2007: 46).

| Element examined | Item searched | José Peñín Subcorpus |
|---|---|---|
| Strict superlatives | *Mejor* | 0 |
| | *Uno de los más* | 0 |
| | Total | 0 |
| Words expressing intensity | *Excelente* | 0 |
| | *Perfecto* | 0 |
| | *Intenso* | 1 |
| | *Intensamente* | 0 |
| | *Ligeramente* | 0 |
| | *Sencillamente* | 0 |
| | *Perfección* | 0 |
| | Total | 1 |
| Different ways of designating a wine | *Personal* | 0 |
| | *Único* | 0 |
| | *Clásico* | 0 |
| | Total | 0 |
| Words denoting the concept of plenty, of weight, of power | *Fuerte* | 0 |
| | *Potente* | 13 |
| | Total | 13 |
| | **Total of all subcorpora** | **14** |

Table 6. Results of the analysis of superlatives in the Spanish subcorpus.

First, in our Spanish corpus, wine was not designated by superlatives. Hence there is one less category for comparison among the subcorpora than in English.

Second, contrary to what was noted in English, the leading authority among Spanish writers of wine tasting notes has a very sober style.

Overall, then, both the English and Spanish subcorpora reveal an opposite trend in the levels of use of superlatives: great use as seen in the Parker corpus in English and little use in the Peñín corpus in Spanish.

### 5.3.2. Figures of speech

Besides exaggeration, wine tasting notes are known for their constant use of figures of speech. Caballero (2007, 2010) points out that imagery is a salient characteristic of wine critics' language – as conspicuously illustrated in tasting notes – and she identifies four categories of imagery in particular: metonymical expressions which allude either to discrete entities or to one of their characteristics (Ripe aromas, apple flavour), (b) similes (wines that taste or smell like a fruit cocktail), (c) terms borrowed from other sensory experiences (wines that smell sweet), and (d) metaphorical language (wines described as fortified, tightly-knit, or broad-shouldered, and qualified as shy, monolithic, or square).

Although she mentions scholars who have dealt with metaphors in wine language (Gluck, 2003; Peynaud, 1987; Amoraritei, 2002; Lehrer, 1983 and 1992), she bemoans the fact that insufficient attention has been paid to figures of speech in wine language. The situation has changed to a large degree since 2007, with many more researchers including Caballero herself, Suarez-Toste, Wipf and others discussing this topic in general, and in particular metaphors, often using tasting notes as a corpus. The general consensus is that tasting notes are full of imagery because of the shortage of terms available to articulate smell and taste experiences.

While a quick read through a few tasting notes is enough to convince you of the importance of imagery in them, there are a few specific questions that are of interest to us: a) does imagery dominate equally tasting notes in both languages in our corpus? b) besides metaphor, which is the main figure of speech in tasting notes, what other figures of speech occur often enough to be worthy of attention? and c) is the imagery restricted to the description of the wine or is it also found in the presentation of the critic's reaction to the wine?

To see whether imagery dominates equally in tasting notes in both English and Spanish, we looked at our English and Spanish corpora and identified the number of tasting notes containing figures of speech (as opposed to those that did not) (Tables 7 and 8 below).

| EN | Total no. of tasting notes | No. of tasting notes containing figures of speech | No. of tasting notes not containing figures of speech |
|---|---|---|---|
| Robert Parker subcorpus | 150 | 135 (90%) | 15 (10%) |

Table 7. Figures of Speech in the English corpus.

The above table reveals that figures of speech abound in Parker subcorpus. In some cases, a tasting note may have only one clear-cut figure of speech, as in the following example, where the metaphorical schema "Wine is a person" is evident, with the focus on age:

(5) Young and fruit-forward, the 2007 Cabernet opens with aromas of dark berry fruit. Soft and supple on the palate, its bittersweet chocolate, vanilla toastiness and olive notes enhance the bouquet. Its rich, long finish adds complexity and depth to the wine. Its youth is evident when tasting in comparison to its predecessors. Like all the Cabernets, the 2007 has the structure and the richness to allow it to develop well for at least an additional ten years.

In other tasting notes, several different figures of speech are found one after the other in the same tasting note, where the nose is described using a simile (like breaking a perfume bottle), the aromas are personified by the word "jumping", and the wine is presented as a woman, full-bodied but not elegant:

(6) This wine had the most pronounced nose of the evening. But was it too much of a good thing? The nose was the equivalent of breaking a bottle of perfume in the sink. The aromas kept jumping from the glass into my nose with pepper, prunes, Armangnac, creamy black fruit & a touch of vanilla ice cream. To say the wine was very full bodied and dense is an understatement! The flavors were very deep and the alcohol was not over the top at 13.5%. The mouth relished the deep black fruit interspersed with jammy black & red fruit. This tannic finish lasted close to 50 seconds. Not what I'd call elegant, but I get the quality.

But whether there is a single figure of speech or several in the tasting notes, the majority of English tasting notes (90% in the English subcorpus) contain them. There is therefore little doubt that figures of speech are a salient feature of English tasting notes.

The situation is a bit different in Spanish, where the Peñín corpus reveals very few figures of speech, as only 3 of its tasting notes (6% of the total) contain them.

| EN | Total no. of tasting notes | No. of tasting notes containing figures of speech | No. of tasting notes not containing figures of speech |
|---|---|---|---|
| José Peñín subcorpus | 150 | 3 (6%) | 147 (94%) |

Table 8. Figures of speech in the Spanish corpus.

We have already indicated that Peñín uses a very sober style in his tasting notes, which could explain the lack of imagery in them. His notes are also very brief and to the point, which further limits the use of imagery.

However, interestingly enough, our Spanish corpus, in contrast to the English corpus, did not contain a single tasting note with only one figure of speech. And even in English, there are more tasting notes containing several figures of speech than those containing only one. This further confirms the dominance of figures of speech in both English and Spanish tasting notes.

Of all the figures of speech found in our corpus, there is no doubt that metaphor is predominant. But it is by no means the only one found in tasting notes. Presented below (examples 7 to 16) is a sample of other figures of speech taken from the English and the Spanish corpora.

Parker corpus:

(7) This is the kind of wine to send chills even up my spine.

(8) This wine is akin to eating candy.

(9) It will take its place in the pantheon of all the great La Mission Haut-Brions ever made.

(10) A skyscraper-like mouthfeel.

(11) All those qualities wrap delicately over the mouth with lasting intensity.

(12) It did not float my boat.

(13) I am positive the wine and I will both be better off if we do not meet again.

The figures of speech were not all easy to categorize, as some could be interpreted in more than one way. However, we can confidently state that, in addition to metaphor, we found metonymy ("mouth" for flavour), simile ("This wine is akin to eating candy") personification ("the wine and I will both be better if we do not meet again") and, above all, idiom[1] "it did not float my boat").

A sample of figures of speech other than metaphor in Spanish is found below:

Peñín subcorpus:

(14) Explosión de fruta roja (fruit explosion).

(15) Con recorrido (long finish).

(16) Boca fresco (fresh palate).

In Spanish, there are examples of metonymy (boca) and personification ("Con recorrido"), as in English, but none of simile or idiom. The diversity of figures of speech is therefore much less in Spanish than in English.

Finally, in the English corpus, but not in the Spanish one, figures of speech are used not only to describe the wine (given the acknowledged lack of terms available for this purpose), but also to describe the critic's overall reaction to the wine. One wine did not "float" Parker's boat; another "sent chills up" his "spine"; in more than one note, Parker was "knocked out" by the wine! In other words, figures of speech are not only used in tasting notes to compensate for a lack of distinct wine-tasting terms, but have become the signature of tasting notes. This is the case in English but not so in Spanish.

### 5.3.3. Personal interventions

Wine tasting notes, which constitute the verbal translation of the experience of wine tasting, have two different functions: descriptive on the one hand, and evaluative on the other. And, according to Lehrer (2009: 7), "the evaluative dimension is important and permeates every other dimension, including the descriptive ones." But judging wines is by its nature subjective, as has been clearly shown by a series of experiments conducted by Robert

Hodgson at the California State Fair wine competition[2] (David Derbyshire, "Wine-tasting: It's junk science", *The Observer,* 23 June 2013. http://www.theguardian.com/lifeandstyle/2013/jun/23/wine-tasting-junk-science-analysis).

The subjectivity shows up in wine tasting notes in at least two different ways: first, certain wine critics feel it incumbent on them to use the first person, to put their personal stamp on the note and the judgement it contains; second, certain tasting notes contain what are clearly personal opinions, but without the use of the first person. Both constitute what we have termed personal interventions.

The first person covers both the first person singular (*I,* etc.) and the first person plural (*we,* etc.). It should be noted that the impression created by the use of the first person singular is somewhat different from what results from the use of the first person plural. The first person plural aligns the writer with the winery or the wine producer, and thus involves the writer less intimately with the wine evaluation, although more so than if no personal pronouns were used. The first person singular is an overt indication of the writer's personal feelings about the wine being described.

It is easy to identify the use of the first person in English notes by searching for *I, my, me, mine* (1st person singular markers) on the one hand, and *we our, us, and ours* (1st person plural markers) on the other. Given the omission of subject pronouns in Spanish, object pronouns, possessives, and verb forms had to be examined to determine first person reference in this language. Presented below (tables 9 and 10) are tables showing first person singular and first person plural use in the English and Spanish corpora, followed by a few concrete examples.

| EN | Total no. of tasting notes | No. of tasting notes containing 1st person singular reference | No. of tasting notes not containing 1st person singular reference |
|---|---|---|---|
| Robert Parker subcorpus | 150 | 42 (28%) | 108 (72%) |

Table 9. Personal interventions in the English corpus.

(17) Not what I'd call elegant, but I get the quality.

The Parker corpus contains notes written by individual wine critics, who are speaking in their own name; it is therefore logical for these notes to have far more 1st person singular reference and almost no 1st person plural reference.

The Spanish corpus reveals, once again, the opposite trend.

| EN | Total no. of tasting notes | No. of tasting notes containing 1st person singular reference | No. of tasting notes not containing 1st person singular reference |
|---|---|---|---|
| Robert Parker subcorpus | 150 | 0 | 0 |

Table 10. Personal interventions in the Spanish corpus.

The comments made above with reference to first person reference in the Parker corpus contrasts the Peñín corpus, which contains no first person reference at all; this lack of first person reference is, however, in keeping with Peñín's sober and detached style.

Personal interventions are revealed more often than not by the use of the first person. However, occasionally a comment is made which is clearly a subjective opinion, but which is not prefaced by "I feel that" or an equivalent expression. An example of such comments is found below:

(18) Great now, this is going to be insane later.

### 5.3.4. Address to the reader

The more personal style, marked by the personal interventions discussed above, is heightened in some cases by the writer directly addressing the reader. This is done either by the use of the second person (you/your/yours) or by the use of the imperative ("Enjoy it moderately chilled"). We will now examine to what extent each of these features figure in the English and Spanish corpora.

Surprisingly, the analysis of the English and Spanish corpora shows the same trend: neither Parker nor Peñín addresses the reader in their notes. It would seem that neither of these two authorities feels the need to establish a personal contact with the reader through direct address.

Another way to address the reader is through the use of imperatives to present the writer's recommendations. Instead of saying "This wine should not be drunk for five years" or "I recommend that this wine not be drunk for five years", the writer says "Do not drink this wine for five years". There are instances of such imperatives in all three English corpus:

(19) Drink: 2010-2040.

However, they are not frequent in the English corpus, and such imperatives are not found at all in the Spanish corpus since Peñín's tasting notes are verbless.

### 5.3.5. Conversational style

Personal interventions by the writer and the address to the readers, both discussed in the preceding sections, could be considered part of the conversational style that seems to characterize a number of tasting notes. However, in this section we will examine some linguistic features characteristic of spoken language, informal language and familiar language, which are all found in conversational style.

For English, the features analyzed are taken from those presented by Leech & Svartvik (1975: 28-31; 2002: 9-18, 30-34) as language variety indicators.

Unlinked clauses:

In informal speech, two neighbouring clauses or sentences may be grammatically unlinked, leaving the connection between them implicit and to be inferred by the reader. (Leech & Svartvik, 2002: 195) This phenomenon is seen in all three English subcorpora, as the following examples show:

(20) Subtle and racy, with lemon rind, vanilla cream and dried pineapple. Very spicy and intense. Full-bodied, with great length and flavor. Electrified yet refined, with medium sweetness and a wonderful finish. I love the class of this, and the length. Has afterburners.

Contracted verb forms:

Some English auxiliary verbs have contracted forms (I'm instead of I am) and the use of the contractions is common in spoken and informal English. (Leech & Svartvik, 2002: 253)

(21) Frankly, I could have drunk the entire barrel sample if it hadn't been my first appointment of the day (at 8:15 a.m.)!

(22) We haven't seen a Clinet this good since the late Jean-Michel Arcaute's stunning duo of 1989 and 1990.

Comment clause in end-position:

Comment clauses are so called because they do not add much to the information in a sentence, but comment on its truth, the manner of saying

it, or the attitude. They are only loosely related to the rest of the main clause they belong to and are marked off from it in written English by commas. While they can occur in front-, mid- and end-positions in the sentence, the end-position is mainly restricted to informal speech. (Leech and Svartvik, 1975: 217).

There was no example of comment clauses in the English corpus.

Ellipsis of that between clauses:

While Leech and Svartvik talk primarily about the omission of *that* before a nominal clause in informal use (*I knew he was wrong*) (Leech and Svartvik, 1975: 249), we have extended this to cover the omission of *that* before other types of clauses as well:

(23) La Mission Haut-Brion has made so many great wines over the last 100 years, it would be stupid to say the 2009 somehow exceeds this estate's great classics.

Emotive emphasis:

Leech and Svartvik (2002: 159-163) indicate a number of ways emotion is emphasized in speech. Some of these ways are discernible in some of the English corpus.

Exclamations:

(24) Bravo!

(25) It's a beauty!

Emphatic so and such:

(26) [T]his precocious Pomerol was tasting unusually well for such a young barrel sample.

Intensifying adverbs and modifiers: (discussed in some detail above).

(27) The 2009 Ausone, was produced at probably twice the yields of the absolutely remarkable 2008.

In addition to these grammatical elements indicative of a conversational style, the tasting notes contain lexical elements that are informal in nature. Some examples are provided below:

(28) [L]oads of black berry and black currant fruit (instead of a large quantity of).

(29) [I]s blend of 65% Cabernet Sauvignon and the rest primarily Merlot with a dollop of Cabernet Franc has a whopping 14.5% alcohol (a dollop of instead of a small amount of; whopping instead of huge).

In summary, there are signs of a conversational style in some tasting notes.

Conversational style in the Spanish subcorpora was examined in light of the features of colloquial language in Spanish identified in *Nueva gramática de la lengua española* (2009-2011), and in *Gramática descriptiva de la lengua española* (Carreter: 1999). Those features found in the corpus include the following:

Unusual word order: No sample found.

Verbless sentences: This covers the lack of a verb or the lack of a finite verb in the main clause:

(30) Color cereza, borde violáceo (cherry red in color, purple rim).

Repetition of words to show intensity: No sample found

Use of abbreviations: No sample found

Use of general words: The use of somewhat vague, non-specific words is considered an element of colloquial language.

(31) En boca algo dulzón

(32) Boca fácil

Comment phrase/clause in end-position: This is the Spanish counterpart of the English comment clause in end-position. No samples found.

While there are some features of conversational style in the Spanish corpus (Verbless sentences and use of several words), the Spanish corpus is far more formal and reflective of written language than the corresponding English corpus.

## 6. Conclusion

The purpose of the paper was to examine the comparability of a comparable corpus from the point of view of the purpose for which it is being used. We

had hypothesized that situational characteristics such as the domain, genre and communicative purpose, time criteria and settings, established by McEnery and Xiao (2007), for the comparison of corpora, might not be valid enough in many cases. We have shown that by analysing other aspects such as format, content and style, in addition to McEnery and Xiao`s, comparability of comparable corpora can be guaranteed.

However, our methodology has also proved that sometimes a given comparable corpus, within a given genre, in a given domain and made up of samples published in the same period of time, shows such low levels of correspondence that only differences can be taken out from that comparison.

Here, in table form (table 11), is a summary of the similarities and differences between the two subcorpora in each language. Instead of specific numbers, we use plus and minus signs to indicate the relative presence or absence of a given feature in a given subcorpus.

| Feature | EN RP | ES JP |
|---|---|---|
| Content: Appearance | + | + |
| Format: Divided | − | + |
| Style: Verbless | − | + |
| Style: Superlatives | + | − |
| Style: Figures of speech | + | − |
| Style: 1st person reference | + | − |
| Style: Address to reader | − | − |
| Style: Conversational style | − | − |

Table 11. Similarities and differences between tasting notes.

As we can see there are identifiable features present in or absent from one subcorpus compared to the other to the extent that even though the content is the same, the style is opposite in most of the cases. That statement leads us to conclude that even though some other criteria seem to be necessary when setting up a comparable corpus, sometimes comparison of the same genres in two different languages brings out more differences than similarities. Genres can be so different interlinguistically that no similarities can be identified. This answers our research questions: to what extent can comparable corpora be compared? This depends mainly on the purpose of the corpus and the research it is being set up for On the other hand, our study has shown that certain genres cannot be compared interlinguistically, if the purpose of the comparable corpora is to look for similarities.

However, our conclusions do not claim that only identical samples can be compared but that other aspects should be take into account (such as the purpose of the corpus and the research) when using comparable corpora.

An interesting follow-up study to this one would be to set up a general corpus of wine tasting notes in English and in Spanish (without identifying the origin of the notes) and see if different types of tasting notes could be identified by analysis of the same features we have examined. In other words, one could try to place tasting notes into at least two different categories – wine critics and notes from wineries or notes from wine journals – on the basis of the features they contain.

Article history:
Received 28 July 2018
Received in revised form 22 July 2019
Accepted 30 April 2020

## References

Amoraritei, L. (2002). "La métaphore en oenologie". *Metaphorik.de* 3: 1–12.

Asimov, E. (2014). "A wine critic's realm isn't a democracy". URL: http://www.nytimes.com/2014/04/23/dining/a-wine-critics-realm-isnt-a-democracy.html? [15/07/2018].

Caballero, R. (2007). "Manner-of-motion verbs in wine description". *Journal of Pragmatics* 39: 2095–2114.

Caballero, R. (2009). "Cutting across the senses: Imagery in winespeak and audiovisual promotion" in C. Forceville & E. Urios-Aparisi (eds.), *Multimodal Metaphor*, 73-94. Berlin and New York: Mouton de Gruyter.

Caballero, R. (2017). "From the glass though the nose and the mouth", *Terminology* 23(1): 66-88.

Caballero, R & E. Suarez-Toste (2008). "Translating the senses. Teaching the metaphors in winespeak" in F. Boers and S. Lindstromberg (eds.), *Cognitive Linguistic Approaches to Teaching Vocabulary and Phraseology*, 241–260. Berlin and New York: Mouton de Gruyter.

Crystal, D. (1987). "The structure of language" in D. Crystal (ed.), *The Cambridge Encyclopaedia of Language*, 15–21. Cambridge: C.U.P.

Derbyshire, D. (2013). "Wine-tasting. It's junk science". URL: http://www.theguardian.com/life-andstyle/2013/jun/23/wine-tasting-junk-science-analysis. [09/01/18].

Eagles. URL: http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html [02/07/18].

Gluck, M. (2003). "Wine language. Useful idiom or idiot-speak?" in J. Aitchison & D.M. Lewis (ed.), *New Media Language*, 107–115. London: Routledge. http://www.wsj.com

Lázaro Carreter, F. (1999). *Gramática descriptiva de la lengua española*. Madrid: Real Academia de la Lengua.

Leech, D. & J. Svartvik. (1975). *A Communicative English Grammar*. London: Longman.

Leech, D. & J. Svartvik J. (2002). *A Communicative English Grammar*. London: Routledge.

Lehrer, A. (1983/2009). *Wine & Conversation*. New York: Oxford University Press.

Li, B. & E. Gaussier (2010). "Improving corpus comparability for bilingual lexicon extraction from comparable corpora" in: S. Acedanski, & A. Przepiórkwski (eds.), *Proceedings COLING'10,* 644-652. Stroudsburg: Association for Computational Linguistics.

Li, B. (2012). *Measuring and Improving Comparable Corpus Quality*. Unpublished Doctoral Dissertation. Université de Grenoble.

López-Arroyo, B. & R.P. Roberts (2014). "English and Spanish descriptors in wine tasting terminology". *Terminology* 20(1): 125-149.

López-Arroyo, B. & R.P. Roberts (2015). "Unusual structures in wine tasting notes: An English-Spanish contrastive analysis". *Languages in Contrast* 15(2): 162-181.

López-Arroyo, B. & Roberts, R.P. (2016). "Differences in wine tasting notes in English and Spanish". *Babel* 62(3): 370-401.

López-Arroyo, B. & R.P. Roberts (2017). "Genre and register in comparable corpora: An English/Spanish contrastive analysis". *Meta* 62(1): 114-136.

Maia, B. (2003). "What are comparable corpora?" in S. Neumann & S. Hansen-Schirra (eds.), *Proceedings of the Corpus Linguistics Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives*, 27-34. Lancaster: Lancaster University.

McEnery, A. & Z. Xiao (2007). "Parallel and comparable corpora. What are they up to?" in G. James & G. Anderman (eds.), *Incorporating Corpora: Translation and the Linguist*, 1-35. Clevendon: Multilingual Matters.

Ni, X., Sun, J.-T., Hu, J., & Chen, Z. (2009). "Mining Multilingual Topics from Wikipedia". In J. Quemada & G. León (eds.), *WWW 09'. Proceedings of the 18th International Conference on World Wide Web*, 1155–1156. ACM: New York.

Nordquist, R. (2014). "Top 20 figures of speech". URL: http://grammar.about.com/od/ rhetoricstyle/a/20figures.htm. [22/05/2018].

Peñín, J. 2009. *Guía Peñín de los vinos de España*. Madrid: Grupo Peñín.

Peynaud, E. (1987). *The Taste of Wine: Art and Science of Wine Appreciation*. San Francisco: The Wine Appreciation Guild.

Real Academia de la Lengua Española. 2009. *Nueva gramática de la lengua española*. URL: http://www.rae.es/recursos/gramatica/nueva-gramatica [06/05/2018].

Sharoff, S. (2013). "Measuring the distance between comparable corpora between languages" in S. Sharoff, R. Rapp, P. Zweigenbaum & P. Fung (eds.), *Building and Using Comparable Corpora*, 113-130. Berlin: Springer.

Tobley-Martínez, T. (2007). "The Lone Wolf: A conversation with wine critic Robert Parker". *The Banner* URL: http://www.naplesnews.com/community/bonita-banner/wine_festival_lone_wolf [11/06/18].

Wipf, B. (2010). *Wine Writing Meets MIPVU: Linguistic Metaphor Identification of Tasting Notes*. MA Thesis. Unpublished. Amsterdam: Vrije Universiteit.

**Belén López Arroyo** is an Associate Professor in ESP at the University of Valladolid (Spain). She currently teaches ESP and translation and Corpus Linguistics in the English Studies Degree. Her research interests include Terminology, Contrastive analysis and Translation. She is author of several articles and books related to contrastive analysis of scientific and professional genres and its implication for translation. In the ACTRES team she is in charge of the Rhetoric of Expert-to-Expert Discourse (in different areas) as well as of terminology and its applications for developing writing aids in English for Spaniards. Along with other members of the ACTRES group she is the author of semi-automatic writing generators aids in the field of Oenology.

## NOTES

[1] There is some disagreement about whether idioms are figures of speech or not. For instance, idioms are not included in the list of top figures of speech provided by David Crystal (1987: 70) or Richard Nordquist ("Top 20 Figures of Speech" http://grammar.about.com/od/rhetoricstyle/a/20figures.htm).

However, given that, essentially, a figure of speech is any unit of speech that cannot be properly understood with a literal interpretation, since figurative language is used and given that an idiom fits that description entirely, we have included idioms as a category of figure of speech.

[2] Each panel of four judges was presented with their usual "flight" of samples to sniff, sip and slurp. But some wines were presented to the panel three times, poured from the same bottle each time. The results were compiled and analysed to see whether wine testing really is scientific. The first experiment took place in 2005. The last was in Sacramento earlier this month. Hodgson's findings have stunned the wine industry. Over the years he has shown again and again that even trained, professional palates are terrible at judging wine. Results from the first four years of the experiment, published in the *Journal of Wine Economics*, showed a typical judge's scores varied by plus or minus four points over the three blind tastings. A wine deemed to be a good 90 would be rated as an acceptable 86 by the same judge minutes later and then an excellent 94.

[3] In the case of this category, which includes several individual items that have been analyzed, we have extrapolated the results.

[4] In the case of this category, which includes several individual features that have been analyzed but not quantified, we have made an overall judgement.