Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 52 – 57

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Patterns, fixedness and variability: using PoS-grams to find phraseologies in the language of travel journalism

David Brett, Antonio Pinna*

Dipartimento di Scienze Umanistiche e Sociali, Università degli Studi di Sassari, via Roma 151, Sassari 07100, Italy

Abstract

As considerable attention has been paid in recent years to variability within fixed sequences (e.g. Biber, 2009; and Gray and Biber, 2013), this paper describes the use of a Corpus Linguistics technique, the Part-of-Speech-gram (usually abbreviated to PoS-gram), that allows potential variability across all slots, and is extremely effective for the discovery of phraseologies that might otherwise remain hidden. A PoS-gram is a string of Part-of-Speech categories (Stubbs, 2007) the tokens of which are strings of words that have been annotated with these PoS tags. Hence, in each slot of the PoS-gram, any word can occur as long as it belongs to the PoS category of that particular position. Despite the vast potential of this technique, it has up to now been largely underused.

This paper will illustrate the utility of PoS-grams by way of analysis of a 450,000 token corpus composed of travel journalism texts from the BBC website. The PoS-grams extracted are compared with a database of PoS-grams obtained from the 100M token BNC. While a large number were found to be statistically significant, in-depth analysis was conducted on PoS-grams containing the inflected superlative adjective form AJS, a feature previously recognized as being central to tourism/travel-writing though without reference to corpus-based techniques (e.g. Dann, 1996).

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: Phraseology; PoS-grams; tourism/travel writing;

* Corresponding author. Tel.: +39-079-229612; fax: +39-079-228211.
E-mail address: dedalo@uniss.it

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.418

1. Introduction

In the present study, attention will be focused on an oft-noted characteristic of the tourism/travel-writing genre, that of Euphoria. Dann (1996) describes how this genre almost invariably presents destinations and the activities to be carried out through rose-tinted glasses, glossing over any negative, unpleasant, or even mundane, aspects that may be encountered by the potential traveller/tourist. Furthermore, there is a tendency to emphasise only the most endearing, fascinating and spectacular facets of the proposed trip or destination. Prior to Dann, Febas Borra (1978) in a study of English language pamphlets issued in Spain writes: “there exists in the language of tourism an obsession with breaking records, of heading up a non-existent classification without any sort of justification”. Dann himself cites Cazes (1976), who observes that such texts present “a 'euphoric global vision', a 'verbal incontinence' in which the superlative is 'de rigueur'”.

It is precisely the latter element that has been selected for analysis in this work. Febas Borra (1978) and Cazes (1976) were writing in a period before corpus linguistics techniques became mainstream. Dann (1996) too does not purport to adopt this methodology. Their claim that the use of the superlative in the tourism/travel-writing genre is pervasive, therefore, requires the type of corroboration that corpus linguistics techniques can offer. Using corpus tools that are readily available today, we can substantiate this assertion by identifying and quantifying superlative forms. Moreover, a comparison with occurrences in other genres can indicate whether the use of the superlative is indeed specific to travel-writing.

However, there has been very little research in Corpus Linguistics concerning the phraseology of this particular construction. For example, Manca (2008) carries out a comparative study of the phraseology of qualifying adjectives in a corpus of British farmhouse holidays and Italian *agriturismi*. However, despite highlighting some intriguing differences in usage in the two comparable corpora, the study focuses solely on the base form of adjectives, and no specific reference is made to either the comparative or the superlative form.

The present works aims to contribute to our knowledge of travel writing by exploring what is purported to be one of its defining characteristics: the use of the superlative. Particular attention will be paid to the typical phraseologies of superlatives in the tourism/travel writing genre.

Essentially, corpus linguistics seeks to identify patterns that are elicited from words and strings of word forms and those studying phraseology using corpus linguistics techniques have adopted a number of strategies to elicit patterns from such strings. Often the same strategy goes by different names (see Stubbs, 2007; Greaves and Warren, 2010). The following is a list of the most common definitions of multi-word sequences:

1. *n*-grams (also known as 'lexical bundles'): uninterrupted strings of word forms, in which the order of the tokens is fixed. The letter *n* indicates the number of tokens composing the string, so we may speak of bigrams, trigrams etc. Biber (2009) proposes a length of 4 as being ideal, in that greater lengths tend to yield far fewer results, while shorter lengths often do not allow identification of the function of the string. In fact, 4 is the value of *n* used in a number of studies (such as Forchini and Murphy, 2008)
2. skip-grams: these are similar to *n*-grams in that they are of a specified length, however, they allow for one or more variable slots, e.g. the skip-gram type *in the * of* would include tokens such as *in the middle of, in the centre of,* and *in the end of* (Biber, 2009)
3. conc-grams: “sets of words that co-occur regardless of constituency variation (e.g. AB and A# B), positional variation (e.g. AB and BA), or both” (Cheng *et al*, 2009).

However, this study outlines the use of a technique that can be considered an even more flexible query type than that of the *n*-gram, skip-gram or conc-gram. The Part-of-Speech-gram (usually abbreviated to PoS-gram) is a string of Part-of-Speech categories (Stubbs, 2007), the tokens of which are strings of words that have been annotated with these PoS tags. Hence, in each slot of the PoS-gram, any word can occur as long as it belongs to the PoS category of that particular position. By casting a considerably looser net than that of the *n*-gram and the skip-gram, PoS-grams are potentially very effective in the discovery of relatively long sequences that fly below the statistical radar of the former techniques. To the best of our knowledge, this study constitutes the first application of the PoS-gram technique to the study of phraseology in the language of Tourism/Travel writing.

2. Corpus Data and Methods

This paper investigates recurrent syntactic and lexical patterns within a corpus of travel journalism. In July 2011, 572 articles, amounting to c. 450,000 tokens, were downloaded from the Travel section of the BBC website [http://www.bbc.com/travel]. This corpus is called the BBC Travel Corpus (henceforth BTC). The provenance of the authors of the text is somewhat heterogeneous: many of the texts were written in collaboration with *Lonely Planet Guides*, or taken directly from these publications. In fact, for 110 articles the only indication of authorship is *Lonely Planet*. Besides, a total of 181 authors are named, some with affiliations (such as *Lonely Planet* or *BBC Olive Magazine*), others without. Excluding the cases where only *Lonely Planet* is mentioned, the average number of articles per author is 2.6. The articles describe a large number of destinations (63) all over the world, the most written about being Italy (35), USA (33) and Great Britain (32). Similarly, a large array of subjects are featured, the most popular being Food and Drink (81), Nature and Outdoors (59) and Cultural Activities (52).

The texts were then tagged for Part-of-Speech (PoS) using the online CLAWS tool [http://ucrel.lancs.ac.uk/claws/]. The tagset used was C5 and the output style was set to Vertical.

The reference corpus chosen was the 100 million word British National Corpus (henceforth BNC). Although larger corpora are now available, the BNC was deemed sufficiently large and varied enough to be representative of general spoken and written usage. Furthermore, it is composed of samples of British English, thus being diatopically coherent with the BTC, and is tagged with the same tagset used for the BTC, thus allowing direct comparison.

Tailormade perl scripts were then used to form PoS-grams starting from each token of the texts. Initially, the length of the PoS-grams was set to 4, however, this resulted in an unmanageable quantity of data. Thereafter, the length was gradually increased up to 8, at which point the results were too small in number to allow in-depth analysis. Therefore, we settled on 6 as an ideal medium that provides both sufficient data and a long enough span to permit the identification of specific functions.

The PoS-grams obtained were then quantified and compared with a database of PoS-grams retrieved from the 100M token BNC in the following way: the PoS-grams extracted from the BTC with f>=10 were tallied with those from the BNC with f>=200. The chi-square test was then applied to identify those that correlated positively with the former corpus. A large number (755) of PoS-grams were found to be typical of the BTC with a significance of $p<0.001$. Twelve of these contained the AJS (superlative adjective) tag. These were marked for further study along with another two which were found to be significant with $p<0.05$.

3. Results and Discussion

3.1. The frequency of the superlative form

We restricted our study of the superlative form to those adjectives with the inflectional suffix *-est*, or suppletion forms such as *better* and *best*, all annotated with the AJS PoS tag, while superlative forms created with *most*+AJ0 were not considered. On the basis of the data obtained, our first research question, that of whether the superlative form is more frequent in travel-writing than in other genres, can be answered in the affirmative. From a simple count of the PoS categories of each token in the corpus, we found that in comparison to data concerning the frequencies in the LSWE (from Biber *et al.,*1999), the superlative (AJS) is ten times more frequent than the register with the highest frequencies in the LSWE, that of news, and fifteen times that of the average of the LSWE (Table 1). Hence we may conclude that, even though travel journalism would generally be considered a type of newspaper language, these results suggest that it may have specific and distinctive formal properties, which in turn may be indicative of it having a different function to mainstream newspaper language.

Table 1. The proportions of tokens annotated with the AJS tag in the different sections of the LSWE and the BTC.

	AJS	Total tokens	%
LSWE Conversation	500	3929500	0.01
LSWE Fiction	700	4980000	0.01
LSWE News	1400	5432800	0.03
LSWE Academic	800	5331800	0.02
Total LSWE	3400	19674100	0.02
BTC	1359	450000	0.3

3.2 PoS-gram analysis

Table 2 lists the 14 PoS-grams containing superlatives that were found to correlate with the BTC in a statistically significant manner.

Table 2. The statistically significant PoS-grams in the BTC that contain AJS tag.

	POS-gram					p	Example
		TC	NC	^{2}	Value		
	AT0 AJS NN1 PRP AT0					p	the highest peak in the country
NN1		3	65	3.39	<0,001		
	CRD PRF AT0 NN1 POS					p	one of the world's largest
AJS		3	98	0.81	<0,001		
	CRD PRF AT0 AJS NN2					p	one of the best kitchens in
PRP		8	30	5.7	<0,001		
	PRF AT0 AJS NN2 PRP					p	of the longest runs in the
AT0		8	62	7.53	<0,001		
	AT0 AJS NN1 NN1 PRP					p	the largest wildlife refuge in the
AT0		1	22	4.57	<0,001		
	AT0 AJS NN2 PRP AT0					p	the longest runs in the land
NN1		8	44	1.72	<0,001		
	PRF AT0 AJS NN2 PRP					p	of the greatest adventures on earth
NN1		0	03	9.64	<0,001		
	PRF AT0 NN1 POS AJS					p	of the world's largest city
NN1		4	09	5.95	<0,001		
	DT0 PRF AT0 NN1 POS					p	some of the world 's best
AJS		3	9	4.12	<0,001		
	PRF AT0 NN1 POS AJS					p	of the world's richest cities
0 NN2		4	38	2.99	<0,001		
	AT0 NN1 POS AJS NN1					p	the world's largest city parks
1 NN2		2	8	1.99	<0,001		
	CRD PRF AT0 AJS NN2					p	one of the best ways to
2 TO0		3	0	1.38	<0,001		
	PRF AT0 AJS NN2 TO0					p	of the best ways to encounter
3 VVI		5	4	.02	<0,05		
	AT0 NN1 POS AJS AJ0					p	the world's largest indoor theme
4 NN1		9	7	.43	<0,05		

What strikes one immediately is the fact that many of these PoS-grams are actually overlapping and appear to be reducible to three essential categories, two of which provide a range or term of comparison, while the third does not. By the term *range* we mean the group from which the specific item is selected and indicated as being that bearing the specific quality to the greatest degree (e.g. *the highest peak in the country* and *the world's largest city parks* as opposed to *of the best ways to encounter*). The results indicate that in travel journalism range is commonly expressed in two ways: using a genitive construction, or a prepositional phrase that post-modifies the head noun.

Examples of the former include PoS-grams 2 and 9 (*one/some of the world's largest*), which may overlap with PoS-grams 10 (*of the world's richest cities*), which is inherently complete, or 8 (*of the world's largest city*), which is inherently incomplete, probably merging in turn with PoS-gram 11 (*the world's largest city parks*) or 14 (*the world's largest indoor theme*), the latter being incomplete, requiring a noun to the right to complete the sequence. In such cases, it is extremely common to use what would be the maximum range in the field of comparison, i.e. *the world*. For instance, in PoS-gram 10, occupants of the third slot in the 24 instances include: *world* (12), *city* (5) and *country* (2). The results for PoS-gram 2 are similar and out of 33 instances, the third slot is occupied by *world* 17 times, followed by *city* (6) and *country* (5). In any case, the nouns in the genitive constructions are almost exclusively of a geographical nature.

The expression of range via prepositional phrases also accounts for a considerable proportion of the significant PoS-grams. Some of these are inherently complete, such as PoS-grams 6 and 7 (*the longest runs in the land* and *of the greatest adventures on earth*, respectively). Others, such as 3, 4 and 5, contain a stub of a prepositional phrase, the noun of which lies beyond the sixth slot. As regards the PoS-grams of this type that are inherently complete (6 and 7), we may observe that there is also a tendency to avail of the maximum range possible. In PoS-gram 6, out of 18 instances, slots 4 to 6 are occupied by *in the world* nine times. A less frequent, but interesting variation of this is what could be considered a marked version: *on the planet*. In this case the author possibly feels that *in the world* is a little overused and opts for an 'alien's eye' view of the situation. A similar phenomenon can be observed in PoS-gram 7. Four out of the ten instances contain the prepositional phrase *on earth*. In one case, the name of the planet is even capitalised, as if to imply that if one seeks something more, the range should be extended to the rest of the solar system, or even beyond!

As noted above, some PoS-grams contain superlatives for which no range is given whatsoever. This is the case of PoS-grams 12 and 13, which clearly share five out of the six slots: *one <u>of the best ways to</u> encounter* (overlapping parts underlined). Examination of the tokens of PoS-gram 13 (Table 3 below) makes it clear that the head noun is followed directly by a *to*-infinitive, the verbs of which are frequently associated with the dative functional role, such as *experience* (3)*, encounter* (1) and *sample* (1). Furthermore, slot 4 shows very little lexical variation and is occupied exclusively by general words for activities and localities: *ways* (8) and *places* (5) with only two exceptions (*islands* and *times*). It is precisely in examples such as this that one can appreciate the importance of the use of PoS-grams in the study of phraseology: *(one) of the best ways to* +VERB is clearly related to both *(one) of the nicest/cheapest ways to* +VERB and *(one) of the best places to* +VERB, however analysis with more traditional corpus linguistics techniques (e.g. *n*-grams) cannot highlight the closeness of the semantic aspects of various tokens, which in turn is a result of their closeness in function.

Table 3. Tokens of PoS-gram 13.

PRF	AT0	AJS	NN2	TO0	VVI
of	the	best	islands	to	cruise
of	the	best	places	to	begin
of	the	best	places	to	get
of	the	best	places	to	go
of	the	best	places	to	reflect
of	the	best	places	to	ride
of	the	best	times	to	stroll
of	the	best	ways	to	encounter
of	the	best	ways	to	experience
of	the	nicest	ways	to	experience
of	the	best	ways	to	experience

of	the	best	ways	to	pass
of	the	best	ways	to	sample
of	the	cheapest	ways	to	ski
of	the	best	ways	to	spend

4. Conclusions

The results of this study not only substantiate previous claims that superlatives are characteristic of tourism/travel-writing, but also demonstrate that this presence is no less than overwhelming: the superlative inflected form is present in the BTC in proportions fifteen times greater than that of a general corpus, the LSWE. We may speculate on the reasons for this: the authors probably make use of the strongest adjective form to attract their readers' attention by way of a claim concerning the extraordinary nature of the locality or product being described. It may also be an attempt to convince the reader that the text he or she is reading will provide better information, suggestions, tips etc. than that offered by other texts dealing with similar topics, hence the high frequency of clusters such as: *the best way to*.

PoS-gram analysis revealed that in comparison to the reference corpus, travel writing makes use of a small series of highly frequent constructions featuring inflectional superlatives. Furthermore, the lexical variation within these constructions is very low. Therefore a methodological point can be made which goes beyond the mere scope of the present study: the use of PoS-grams for the study of phraseology is at present highly underused. The quantitative analysis of strings of PoS categories and their relating tokens casts a looser net over a wider area, allowing us to discover widespread and characteristic patterns that fly below the statistical radar of more traditional and stricter forms of analysis such as *n*-grams, which can only reveal identity and not similarity.

References

Baker, P., Hardie, A., and McEnery, A. (2006). *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow: Longman.

Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics, 14(3),* 275 - 311.

Cazes, G. (1976). *Le tiers-monde vu par le publicités touristiques: une image mystifiante. cahiers du tourisme*, série C 33.

Cheng, W., Greaves C., Sinclair, J., and Warren M. (2009). Uncovering the extent of the phraseological tendency: towards a systematic analysis of Concgrams. *Applied Linguistics*, *30*, 236 – 252.

Dann, G. (1996). *The language of tourism: a sociological perspective*. Wallingford: CABI Publishing.

Febas Borra, J.L. (1978). Semiología del lenguaje turístico. *Estudios Turísticos, 57-58*, 17 - 204.

Forchini, P., and Murphy, A. (2008). N-grams in comparable specialized corpora. *International Journal of Corpus Linguistics*, *13*, 351 - 367.

Gray, B., and Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, *18*, 109 - 135.

Greaves, C. and Warren, M. (2010). What can a corpus tell us about multi-word units? In O'Keeffe, A., and M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 212-226). New York: Routledge.

Manca, E. (2008). From phraseology to culture. Qualifying adjectives in the language of tourism. *International Journal of Corpus Linguistics 13*, 368 - 385.

Stubbs, M. (2007). An example of frequent English phraseology: distributions, structures and functions. In Facchinetti, R. (Ed.), *Corpus linguistics 25 Years on* (pp. 89-105). Amsterdam: Rodopi.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Amsterdam: John Benjamins.