Automatically Assessing Whether a Text Is Clichéd, with Applications to Literary Analysis

Paul Cook
Department of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia
paulcook@unimelb.edu.au

Graeme Hirst
Department of Computer Science
University of Toronto
Toronto, ON, Canada M5S 3G4
gh@cs.toronto.edu

Abstract

Clichés, as trite expressions, are predominantly multiword expressions, but not all MWEs are clichés. We conduct a preliminary examination of the problem of determining how clichéd a text is, taken as a whole, by comparing it to a reference text with respect to the proportion of more-frequent *n*-grams, as measured in an external corpus. We find that more-frequent *n*-grams are over-represented in clichéd text. We apply this finding to the "Eumaeus" episode of James Joyce's novel *Ulysses*, which literary scholars believe to be written in a deliberately clichéd style.

1 Clichés

In the broadest sense a cliché is a tired, overused, unoriginal idea, whether it be in music, in the visual arts, in the plot of a novel or drama, or in the language of literature, journalism, or rhetoric. Here, we are interested only in clichés of linguistic form. Clichés are overused, unoriginal expressions that appear in a context where something more novel might have reasonably been expected, or which masquerade as something more original, more novel, or more creative than they actually are. A cliché is a kind of ersatz novelty or creativity that is, *ipso facto*, unwelcome or deprecated by the reader. Clichés appear to be intuitively recognized by readers, but are difficult to define more formally.

Clichés are predominantly multiword expressions (MWEs) and are closely related to the idea of formulaic language, which for Wray (2002, 2008, summarized in 2009) is a psycholinguistic phenomenon: a formula is stored and retrieved as a single prefabricated unit, without deeper semantic analysis, even if it is made up of meaningful smaller units and regardless of whether it is or isn't semantically transparent. She demonstrates that formulaic language is a heterogeneous phenomenon, encompassing many types of MWEs including fixed expressions (Sag et al., 2002, e.g., *whys and wherefores*), semi-fixed expressions (e.g., *hoist with/by his own petard* 'injured by that with which he would injure others'), and syntactically-flexible expressions (e.g., sb_1 *haul* sb_2 *over the coals* 'reprimand severely', allowing also the passive sb_2 *was hauled over the coals (by* sb_1*)*). Formulaic language can exhibit any of the types of idiomaticity required by Baldwin and Kim (2010) for an expression to be considered an MWE, i.e., lexical (*de rigueur*), syntactic (*time and again*), semantic (*fly off the handle* 'lose one's temper'), pragmatic (*nice to see you*), and statistical idiomaticity (which many of the previous examples also exhibit).

Another theme relating formulaic language to MWEs is that of a common or preferred (though not necessarily invariable) way for native speakers to express an idea, i.e., institutionalization; for example, felicitations to someone having a birthday are usually expressed as *happy birthday* or (largely in British English) *many happy returns* rather than any of the many other semantically similar possibilities (#*merry birthday*; cf. *merry Christmas*).

However, formulaic language, including clichés, goes beyond the typical view of MWEs in that it has a cultural aspect as well as a purely linguistic aspect, as it includes catchphrases and allusions to language in popular culture, such as well-known

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 52–57,
Atlanta, Georgia, 13-14 June 2013.

lines from songs, jokes, advertisements, books, and movies (*curiouser and curiouser* from Lewis Carroll's *Alice's Adventures in Wonderland*; *go ahead, make my day* 'I dare you to attack me or do something bad, for if you do I will take great pleasure in defeating and punishing you' from the 1983 Clint Eastwood movie *Sudden Impact*).

Furthermore, not all formulaic language is clichéd; a weather forecast, for example, has no pretensions of being linguistically creative or original, but it would be a mistake to think of it as clichéd, no matter how formulaic it might be. Conversely, a cliché might not be formulaic from Wray's psycholinguistic perspective — stored and recognized as a single unit — even if its occurrence is at least frequent enough in relevant contexts for it to be recognized as familiar, trite, and unoriginal.

Finally, not all MWEs are clichés. Verb–particle constructions such as *look up* ('seek information in a resource') and *clear out* are common expressions, but aren't unoriginal in the sense of being tired and over-used. Moreover, they are not *attempts* at creativity. On the other hand, clichés are typically MWEs. Some particularly long clichés, however, are more prototypical of proverbs than MWEs (e.g., *the grass is always greener on the other side*). Single words can also be trite and over-used, although this tends to be strongly context dependent.

This paper identifies clichés as an under-studied problem closely related to many issues of interest to the MWE community. We propose a preliminary method for assessing the degree to which a text is clichéd, and then show how such a method can contribute to literary analysis. Specifically, we apply this approach to James Joyce's novel *Ulysses* to offer insight into the ongoing literary debate about the use of clichés in this work.

2 Related work

Little research in computational linguistics has specifically addressed clichés. The most relevant work is that of Smith et al. (2012) who propose a method for identifying clichés in song lyrics, and determining the extent to which a song is clichéd. Their method combines information about rhymes and the df-idf of trigrams (tf-idf, but using document frequency instead of term frequency) in song lyrics. However, this method isn't applicable for our goal of determining how clichéd an arbitrary text is with a focus on literary analysis, because in this case rhyming is not a typical feature of the texts. Moreover, repetition in song lyrics motivated their df-idf score, but this is not a salient feature of the texts we consider.

In his studies of clichés in *Ulysses*, Byrnes (2012) has drawn attention to the concept of the *cliché density* of a text, i.e., the number of clichés per unit of text (e.g., 1000 words). Byrnes manually identified clichés in *Ulysses*, but given a comprehensive cliché lexicon, automatically measuring cliché density appears to be a straightforward application of MWE identification — i.e., determining which tokens in a text are part of an MWE. Although much research on identification has focused on specific kinds of MWEs (Baldwin and Kim, 2010), whereas clichés are a mix of types, simple regular expressions could be used to identify many fixed and semi-fixed clichés. Nevertheless, an appropriate cliché lexicon would be required for this approach. Moreover, because of the relationship between clichés and culture, to be applicable to historical texts, such as for the literary analysis of interest to us, a lexicon for the appropriate time period would be required.

Techniques for MWE extraction could potentially be used to (semi-) automatically build a cliché lexicon. Much work in this area has again focused on specific types of MWEs — e.g., verb–particle constructions (Baldwin, 2005) or verb–noun combinations (Fazly et al., 2009) — but once more the heterogeneity of clichés limits the applicability of such approaches for extracting them. Methods based on strength of association — applied to *n*-grams or words co-occurring through some other relation such as syntactic dependency (see Evert, 2008, for an overview) — could be applied to extract a wider range of MWEs, although here most research has focused on two-word co-occurrences, with considerably less attention paid to longer MWEs. Even if general-purpose MWE extraction were a solved problem, methods would still be required to distinguish between MWEs that are and aren't clichés.

3 Cliché-density of known-clichéd text

Frequency per se is not a necessary or defining criterion of formulaic language. Wray (2002) points out that even in quite large corpora, many undoubted instances of formulaic language occur infrequently or not at all; for example, Moon (1998) found that formulae such as *kick the bucket* and *speak for yourself!* occurred zero times in her 18 million–word representative corpus of English. Nevertheless in a very large corpus we'd expect a formulaic expression to be more frequent than a more-creative expression suitable in the same context. Viewing clichés as a type of formulaic language, we hypothesized that a highly-clichéd text will tend to contain more *n*-grams whose frequency in an external corpus is medium or high than a less-clichéd text of the same size.

We compared a text known to contain many clichés to more-standard text. As a highly-clichéd text we created a document consisting solely of a sample of 1,988 clichés from a website (`clichesite.com`) that collects them.[1] For a reference "standard" text we used the written portion of the British National Corpus (BNC, Burnard, 2000). But because a longer text will tend to contain a greater proportion of low-frequency *n*-gram types (as measured in an external corpus) than a shorter text, it is therefore crucial to our analysis that we compare equal-size texts. We down-sampled our reference text to the same size as our highly-clichéd text, by randomly sampling sentences.

For each 1–5-gram type in each document (i.e., in the sample of clichés and in the sample of sentences from the BNC), we counted its frequency in an external corpus, the Web 1T 5-gram Corpus (Web 1T, Brants and Franz, 2006). Histograms for the frequencies are shown in Figure 1. The *x*-axis is the log of the frequency of the *n*-gram in the corpus, and the *y*-axis is the proportion of *n*-grams that had that frequency. The dark histogram is for the sample from the BNC, and the light histogram is for the clichés; the area where the two histograms overlap is medium grey. For 1-grams, the two histograms are quite similar; hence the following observations are not merely due to simple differences in word frequency. For the 3–5-grams, the light areas show that the clichés contain many more *n*-gram types with medium or high frequency in Web 1T than the sample of sentences from the BNC. For each of the 3–5-grams, the types in the sample of clichés are significantly more frequent than those in the BNC using a Wilcoxon rank sum test ($p \ll 0.001$). The histogram for the 2-grams, included for completeness, is beginning to show the trend observed for the 3–5-grams, but there is no significant difference in mean frequency in this case.

This finding supports our hypothesis that clichéd text contains more higher-frequency *n*-grams than standard text. In light of this finding, in the following section we apply this *n*-gram–based analysis to the study of clichés in *Ulysses*.

4 Assessing cliché-density for literary analysis

Ulysses, by James Joyce, first published in 1922, is generally regarded as one of the greatest English-language novels of the twentieth century. It is divided into 18 episodes written in widely varying styles and genres. For example, some episodes are, or contain, long passages of stream-of-consciousness thought of one of the characters; another is written in catechism-style question-and-answer form; some parts are relatively conventional.

Byrnes (2010, 2012) points out that it has long been recognized that, intuitively, some parts of the novel are written in deliberately formulaic, clichéd language, whereas some other parts use novel, creative language. However, this intuitive impression had not previously been empirically substantiated. Byrnes took the simple step of actually counting the clichés in four episodes of the book and confirmed the intuition. In particular, he found that the "Eumaeus" episode contained many more clichés than the other episodes considered. However, these results are based on a single annotator identifying the clichés — Byrnes himself — working with an informal definition of the concept, and possibly biased by expected outcomes. By automatically and objectively measuring the extent to which "Eumaeus" is clichéd, we can offer further evidence — of a very different type — to this debate.

[1] Because we don't know the coverage of this resource, it would not be appropriate to use it for an MWE-identification approach to measuring cliché-density.

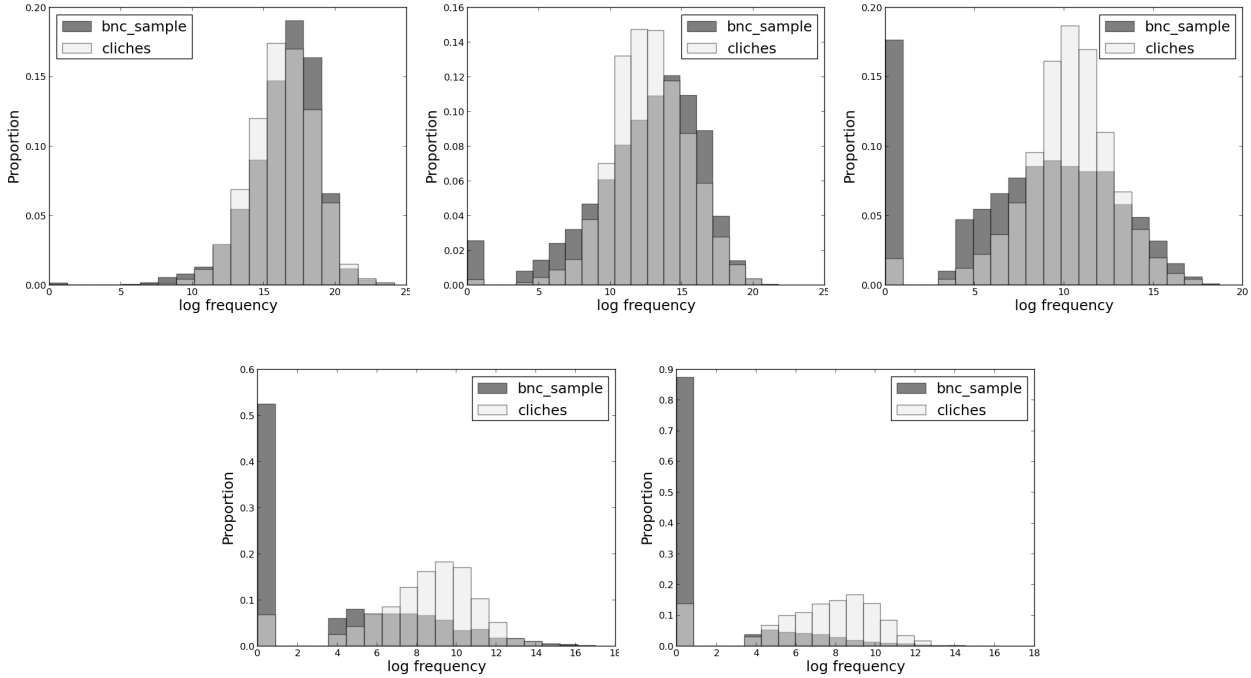

Figure 1: Histograms for the log frequency of *n*-grams in a sample of sentences from the BNC and a collection of known clichés. 1–5-grams are shown from left to right, top to bottom.

We compared "Eumaeus" to a background text consisting of episodes 1–2 and 4–10 of *Ulysses*, which are not thought to be written in a marked style. Because formulaic language could vary over time, we selected an external corpus from the time period leading up to the publication of *Ulysses* — the Google Books NGram Corpus (Michel et al., 2011) for the years 1850–1910 (specifically, the "English 2012" version of this corpus). We down-sampled each episode, by randomly sampling sentences, to the size of the smallest, to ensure that we compared equal-size texts.

Figures 2 and 3 show histograms for the frequencies in the external corpus of the 1–5-grams in "Eumaeus" and in the background episodes. If "Eumaeus" is more-clichéd than the background episodes, then, given our results in Section 3 above, we would expect it to contain more high-frequency higher-order *n*-grams. We indeed observe this in the histograms for the 3- and 4-grams. The differences for each of the 3–5-grams are again significant using Wilcoxon rank sum tests ($p \ll 0.001$ for 3- and 4-grams, $p < 0.005$ for 5-grams), although the effect is less visually striking than in the analysis in Section 3, particularly for the 5-grams. One possible reason for this difference is that in the analysis in Section 3 the known-clichéd text was artificial in the sense that it was a list of expressions, as opposed to natural text.

We further compared the mean frequency of the 3-, 4-, and 5-grams in "Eumaeus" to that of each individual background episode, again down-sampling by randomly sampling sentences, to ensure that equal-size texts are compared. In each case we find that the mean *n*-gram frequency is highest in "Eumaeus". These results are consistent with Byrnes's finding that "Eumaeus" is written in a clichéd style.

5 Conclusions

Clichés are an under-studied problem in computational linguistics that is closely related to issues of interest to the MWE community. In our preliminary analysis, we showed that a highly-clichéd text contains more higher-frequency *n*-gram types than a more-standard text. We then applied this approach to literary analysis, confirming beliefs about the use of clichés in the "Eumaeus" episode of *Ulysses*.

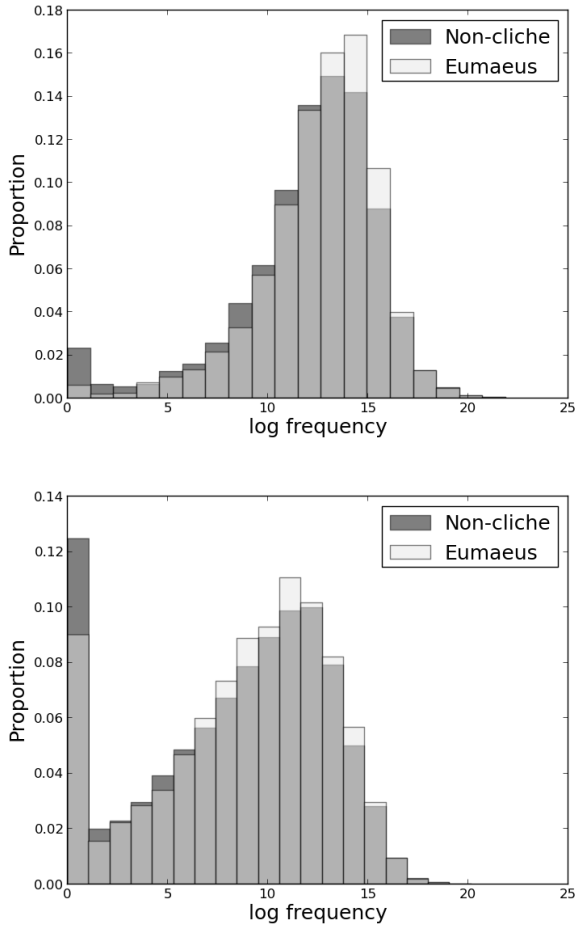

Figure 2: Histograms for the log frequency of *n*-grams in the "Eumaeus" episode of Ulysses and episodes known to be non-clichéd. 1-, and 2-grams are shown on the top and bottom, respectively.

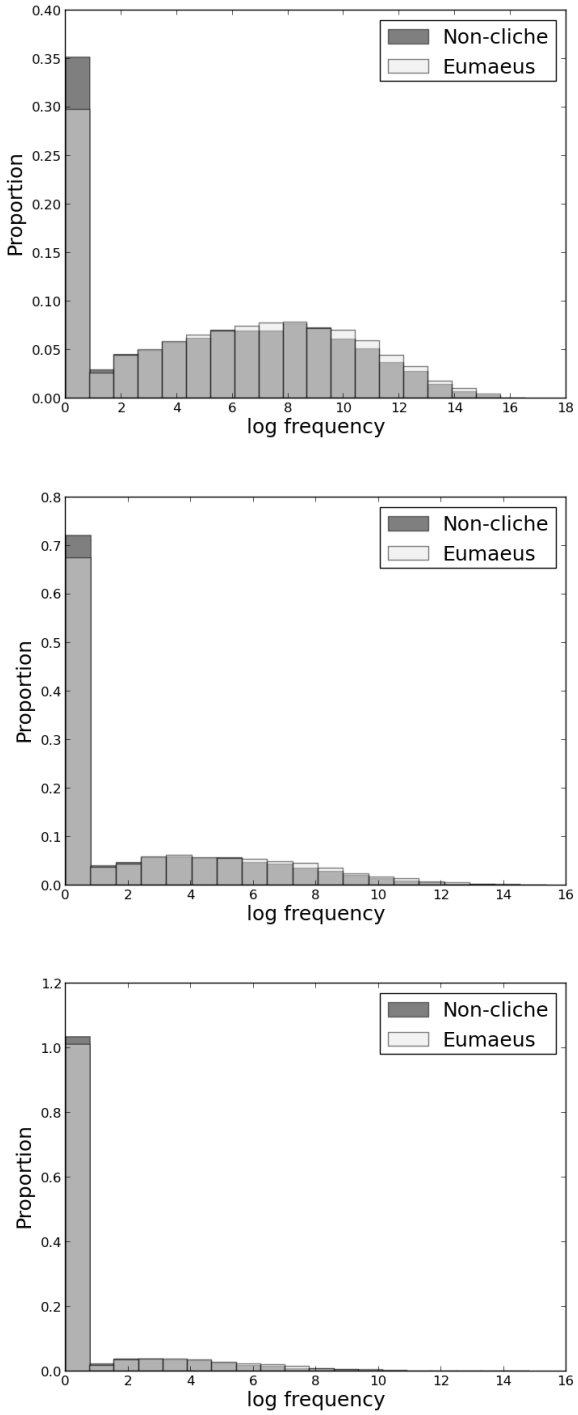

Figure 3: Histograms for the log frequency of *n*-grams in the "Eumaeus" episode of Ulysses and episodes known to be non-clichéd. 3-, 4-, and 5-grams are shown on the top, middle, and bottom, respectively.

Acknowledgments

We thank Timothy Baldwin and Bahar Salehi for their insightful comments on this work. This work was supported financially by the Natural Sciences and Engineering Research Council of Canada.

References

Timothy Baldwin. 2005. The deep lexical acquisition of English verb-particle constructions. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):398–414.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J.

Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 267–292. CRC Press, Boca Raton, USA.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Corpus version 1.1.

Lou Burnard. 2000. *The British National Corpus Users Reference Guide*. Oxford University Computing Services.

Robert Byrnes. 2010. A statistical analysis of the "Eumaeus" phrasemes in James Joyce's *Ulysses*. In *Actes des 10es Journées internationales d'Analyse statistique des Données Textuelles / Proceedings of the 10th International Conference on Textual Data Statistical Analysis*, pages 289–295. Rome, Italy.

Robert Byrnes. 2012. The stylometry of cliché density and character in James Joyce's *Ulysses*. In *Actes des 11es Journées internationales d'Analyse statistique des Données Textuelles / Proceedings of the 11th International Conference on Textual Data Statistical Analysis*, pages 239–246. Liège, Belgium.

Stefan Evert. 2008. Corpora and collocations. In *Corpus Linguistics. An International Handbook*. Article 58. Mouton de Gruyter, Berlin.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Rosamund Moon. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Clarendon Press.

Ivan A. Sag, 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 Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15.

Alex G. Smith, Christopher X. S. Zee, and Alexandra L. Uitdenbogerd. 2012. In your eyes: Identifying clichés in song lyrics. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 88–96. Dunedin, New Zealand.

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press.

Alison Wray. 2008. *Formulaic Language: Pushing the Boundaries*. Oxford University Press.