Pop lyrics and language pedagogy: A corpus-linguistic approach

Valentin Werner
University of Bamberg
valentin.werner@uni-bamberg.de

Maria Lehl
Tonguesten
maria@tonguesten.com

Introduction

Although English pop music lyrics are part of most people's everyday life, to date, they have largely been ignored in corpus linguistic research. This can be deduced from the facts that lyrics rarely form part of standard corpora of English and that the amount of corpus-based research explicitly devoted to this register (e.g. Kreyer & Mukherjee 2007 or Werner 2012) is restricted. In addition, in applied linguistic attempts to exploit them for EFL teaching purposes they have mostly – and despite their high motivational value (see e.g. Syed 2001; Beath 2010; Israel 2013) – been sidelined to the role of "additional" or "light" material, usually found at the end of chapters, and barred from the use for the instruction of "serious" matter, such as aspects of grammar (see Murphey 1990, 1995 for notable exceptions).

We will argue that lyrics should emerge from their shadowy existence as they are a worthy subject both for corpus linguists and for practitioners in EFL for various reasons. With that goal in mind, we will address the topic of pop lyrics from three angles. First, we will present a general overview of linguistic features of lyrics and thus offer a brief stylistic analysis (in terms of locating lyrics in relation to other text types as well as on a written-spoken continuum), also considering learner-related aspects. Second, we will widen the perspective and will consider why the NLP annotation of lyrics is notoriously difficult due to some of their inherent features. It will also be discussed how these issues can be overcome. In the final section, we will address the question of how pop lyrics can be used in language teaching and learning (e.g. in terms of web applications such as Tonguesten's Rebeats[1] platform), taking advantage of the specific opportunities offered by a corpus-based approach.

1 Stylistic analysis

The data on which the analyses are based derive from purpose-built sources. One of them is the *Chart Corpus* (cf. Werner 2012), a 342,202-token corpus (1,128 songs) containing lyrics from songs that were highly successful (i.e. at least among the top five) in the UK and the US in the years 1946 to 2008.

A general quantitative comparison to other text types using the Multidimensional Analysis Tagger (Nini 2014) locates lyrics close to the category "informational interaction". This seems surprising as lyrics typically are viewed as a form of one-to-many communication and thus supposedly lack characteristic interactional features. However, when the individual dimensions of variation (following Biber 1988) are considered in detail, an ambiguous picture emerges. The analysis yields lyrics as an involved text type (Dimension 1), but with non-narrative concerns (Dimension 2), for instance. This ties in with previous research which has shown that lyrics can be viewed as a "particular" genre that (i) cannot unequivocally be assigned to the written or spoken mode and (ii) that is characterized by features associated both with formal and informal usage (Werner 2012: 43).

Learners, who receive their formal English instruction largely with the help of textbooks (which aim at a standard form of the target language) may be unfamiliar with a number of nonstandard features that occur in lyrics. Potential hurdles are contractions (*upon* > *'pon*), as well as other elisions, for instance of auxiliaries or third-person markers, all illustrated in (1).

(1) but she gone and she not comeback me beg
her please 'pon me knees and she still never
stop (Pato Banton: "Come back")

Another case in point are nonstandard pronoun and verb forms as in (2) or (3), which may be used as identity markers or can also be interpreted as devices to indicate the cultural hybridity of the text (e.g. realized through a combination of standard and Creole features).

(2) so me say, we a go hear it on the stereo
(Musical Youth: "Pass the Dutchy")
(3) but me know I'm not a fear to you
(Sean Paul & Blu Cantrell: "Breathe")

The motivation of being able to cope with such nonstandard features as well as with the inherent hybridity of lyrics renders them a challenging but equally stimulating resource for learners. Likewise, reliable NLP annotation of lyrics – with available taggers usually trained on standard forms of a language – is a challenging task for the corpus linguist, as will be shown subsequently.

[1] http://www.rebeats.tv

2 NLP annotation

Part-of-speech (POS) tagging is a gateway step into corpus linguistic research of lyrics. However, training a POS tagger would require the annotation of a sufficiently large training corpus that does not exist up to date. In an exploratory study, six pre-trained tagger models using the Penn tag set were assessed on a 100-song gold standard, compiled from the top ten UK albums of the years 2001 to 2011 (Lehl 2014).

With a focus on testing a broad range of tagging approaches, the HunPos tagger (Halácsy et al. 2007), the Stanford Tagger (Toutanova et al. 2003) and the SVM Tool (Giménez and Màrquez 2004) were selected. All tagger models are trained on the Wall Street Journal (WSJ) corpus. However, some models use online chat conversations, Tweets and other web content as additional training data.

The results show tagger model performances ranging between 90.60% and 93.05% and thus well below state-of-the-art of 97-98% on the WSJ corpus. The best-performing model was the Stanford Tagger model, which is trained on the WSJ corpus enriched by a chat corpus and Tweets.[2]

Knowing that all models are trained on the WSJ corpus among others, the taggers were assessed separately on WSJ-tokens, which had been encountered by all taggers in their WSJ training data (*known tokens*), and on non-WSJ tokens. A qualitative analysis shows that, apart from noise-related tagging errors, many of the inaccuracies on the non-WSJ tokens can be traced back to lyrics-specific phenomena, primarily contractions (see above) and musical tropes (such as *yeah* and *woah*). Tagging errors of this type can easily be avoided by using word lists and regular expressions. However, the low accuracy of taggers on non-WSJ tokens contributes only little to the inferior general performance of taggers on lyrics as compared to the performances of taggers on the WSJ corpus. Even on the known tokens alone the maximum tagging accuracy lies at merely 93.52%. An error analysis using confusion matrices revealed common tagging errors to be standard tag confusions, such as the following:

- VB wrongly tagged as VBP

  “Have/VBP* yourself a merry little Christmas”
  (Michael Bublé: “Have Yourself a Merry Little Christmas)

- VBN wrongly tagged as VBD

  “Have you heard/VBD* the news today”
  (P!nk: “Gone to California”)

This poses the question why these common tagging errors occur more frequently in lyrics. One possible explanation is that sentence boundaries are mostly missing. As a consequence, each lyric line was fed to the taggers as one sentence, which may have given insufficient context for tagging. Other possible explanations are that lyrics contain significantly more occurrences of elisions and non-standard grammar (see above) than the training data of the taggers.

These results suggest that increasing the size of provided context for tagging (e.g. by pairwise binding of consecutive lines) and compiling a sufficiently large training corpus are necessary steps of research to engage in. However, an important point of investigation that has to be undertaken before is the computation of a ceiling tagging performance by an inter-rater agreement. There is reason to believe that tagging ambiguity in lyrics is generally higher as the musically imposed shortness of lines, the frequency of ellipsis constructions, incoherent content, and slang make tagging sometimes challenging even to the human annotator.

3 Integrating corpus linguistics and language learning

The limitations illustrated should not disguise the fact that a lot can already be done with annotated lyrics data, thus also going beyond traditional uses of lyrics in educational contexts. A central field for the application of linguistically annotated lyrics corpora is represented by Computer- and Mobile-Assisted Language Learning (CALL and MALL). There has been a recent trend of language learning gamification, one phenomenon being online language courses that target the use of lyrics and song videos for EFL. Rebeats is one example of such a web-based EFL platform in development, which uses linguistically annotated data and offers one road of how findings from lyrics-related linguistic research can be applied. The main goal of the platform is to automatize the creation of language exercises from lyrics, packing the outcome into an engaging multiple-choice game as exemplified in Figure 1. In this case, POS-tagged lyrics are used to automatically create a verb tense exercise targeting the construction of the present perfect in English. The learner is challenged by multiple choice exercises while the video clip is playing, and receives instant feedback.

Examples such as Rebeats show that an integration of corpus-based findings and application in language learning is possible. In the future, the linguistic community should provide more insights on (i) individual features of “popular” content (see

[2] https://gate.ac.uk/wiki/twitter-postagger.html

also Bértuoli-Dutra 2014), (ii) how to deal with them in NLP, and (iii) how to exploit their full pedagogical potential.

References

Beath, O. 2010. "'I want to be more perfect than others': a case of ESL motivation." Paper presented at the Faculty of Education and IERI HDR Conference, Wollongong, 12 November 2010. Available online at http://ro.uow.edu.au/edupapers/161/

Bértuoli-Dutra, P. 2014. "Multi-dimensional analysis of pop songs." In T. B. Sardinha and M. V. Pinto (eds.) *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber.* Amsterdam: Benjamins. 149-176.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Giménez, J. and Màrquez, L. 2004. "SVMTool: A general POS tagger generator based on Support Vector Machines." In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa and R. Silva (eds.) *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Paris: ELRA. 43-46. Available online at http://www.lrec-conf.org/proceedings/lrec2004/pdf/597.pdf

Halácsy, P., Kornai, A. and Oravecz, C. 2007. "HunPos – an open source trigram tagger." In *45th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Interactive Poster and Demonstration Sessions.* Prague: Association for Computational Linguistics. 209-212. Available online at http://aclweb.org/anthology/P07-2

Israel, H. F. 2013. "Language learning enhanced by music and song." *Literacy Information and Computer Education Journal* 2 (1): 1269-1275.

Kreyer, R. and J. Mukherjee. 2007. "The style of pop song lyrics: a corpus-linguistic pilot study." *Anglia* 125 (1): 31-58.

Lehl, M. 2014. *Stairway to Learner's Heaven: Using Song Lyrics to Build a Resource for Automatic Creation of Language Exercises.* Unpublished Master's thesis, University of Osnabrück.

Murphey, T. 1990. *Song and Music in Language Learning: An Analysis of Pop Song Lyrics and the Use of Song and Music in Teaching English to Speakers of Other Languages*. Frankfurt: Peter Lang.

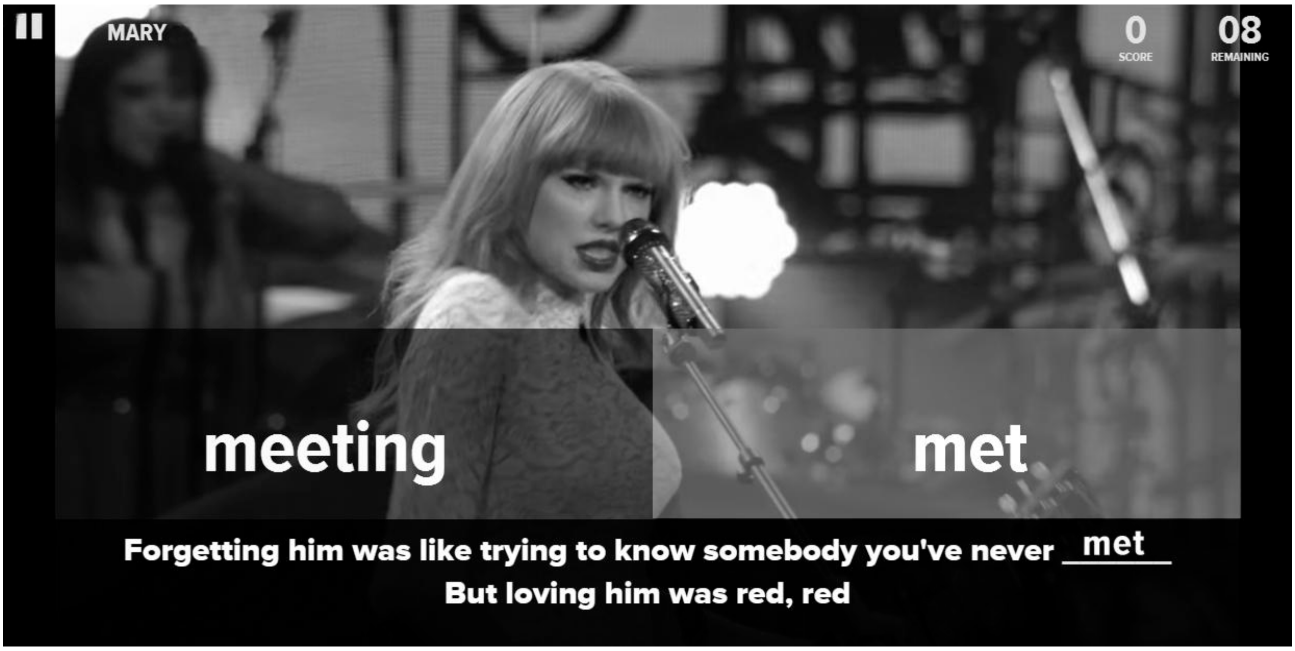
Murphey, T. 1995. *Music and Song.* Oxford: Oxford University Press.

Nini, A. 2014. *Multidimensional Analysis Tagger 1.2.* Available online at http://sites.google.com/site/multidimensionaltagger

Syed, Z. 2001. "Notions of self in foreign language learning: a qualitative analysis." In Z. Dörnyei and R. Schmidt (eds.) *Motivation and Second Language Acquisition*. Honolulu: University of Hawai'i Second Language Teaching and Curriculum Center. 127-148.

Toutanova, K., Klein, D., Manning, C. and Singer, Y. 2003. "Feature-rich part-of-speech tagging with a cyclic dependency network." In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. 173-180. Available online at http://nlp.stanford.edu/pubs/tagging.pdf

Werner, V. 2012. "Love is all around: a corpus-based study of pop music lyrics." *Corpora* 7 (1): 19-50. Available online at http://www.euppublishing.com/doi/pdfplus/10.3366/cor.2012.0016



Forgetting him was like trying to know somebody you've never met
But loving him was red, red

Figure 1: Example of a learning exercise on the online language platform Rebeats