Beefmoves: Dissemination, Diversity, and Dynamics of English Borrowings in a German Hip Hop Forum

Matt Garley
Department of Linguistics
University of Illinois
707 S Mathews Avenue
Urbana, IL 61801, USA
mgarley2@illinois.edu

Julia Hockenmaier
Department of Computer Science
University of Illinois
201 N Goodwin Avenue
Urbana, IL 61801, USA
juliahmr@illinois.edu

Abstract

We investigate how novel English-derived words (anglicisms) are used in a German-language Internet hip hop forum, and what factors contribute to their uptake.

1 Introduction

Because English has established itself as something of a global lingua franca, many languages are currently undergoing a process of introducing new loanwords borrowed from English. However, while the motivations for borrowing are well studied, including e.g. the need to express concepts that do not have corresponding expressions in the recipient language, and the social prestige associated with the other language (Hock and Joseph, 1996), the dynamics of this process are poorly understood. While mainstream political debates often frame borrowing as evidence of cultural or linguistic decline, it is particularly pervasive in youth culture, which is often heavily influenced by North American trends. In many countries around the globe, hip hop fans form communities in which novel, creative uses of English are highly valued (Pennycook, 2007), indicative of group membership, and relatively frequent. We therefore study which factors contribute to the uptake of (hip hop-related) anglicisms in an online community of German hip hop fans over a span of 11 years.

2 The MZEE and Covo corpora

We collected a ∼12.5M word corpus (MZEE) of forum discussions from March 2000 to March 2011 on the German hip hop portal MZEE.com. A manual analysis of 10K words identified 8.2% of the tokens as anglicisms, contrasting with only 1.1% anglicisms in a major German news magazine, the *Spiegel* (Onysko, 2007, p.114). These anglicisms include uninflected English stems (e.g., *battle, rapper, flow*) as well as English stems with English inflection (e.g., *battled, rappers, flows*), English stems with German inflection (e.g., *gebattlet, rappern, flowen* 'battled, rappers, to flow'), and English stems with German derivational affixes (e.g., *battlemässig, rapperische, flowendere* 'battle-related, rapper-like, more flowing'), as well as compounds with one or more English parts (e.g., *battleraporientierter, hiphopgangstaghettorapper, maschinengewehrflow* 'someone oriented towards battle-rap, hip hop-gangsta-ghetto-rapper, machinegun flow'). We also collected a ∼20M word corpus (Covo) of English-language hip hop discussion (May 2003 - November 2011) from forums at ProjectCovo.com.

3 Identification of novel anglicisms

In order to identify novel anglicisms in the MZEE corpus, we have developed a classifier which can identify anglicism candidates, including those which incorporate German material (e.g., *möchtegerngangsterstyle* 'wannabe gangster style'), with very high recall. Since we are not interested in well-established anglicisms (e.g., *Baby*, *OK*), non-English words, or placenames, our goal is quite different from the standard language identification problem, including Alex (2008)'s inclusion classifier, which sought to identify 'foreign words' in general, including internationalisms, homographic

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 135–139,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

Baseline n-gram classifier accuracy for n=						
1	2	3	4	5	6	7
87.54	94.80	97.74	99.35	99.85	99.96	99.98

Figure 1: Accuracy of the baseline classifer on word lists; 10-fold CV; std. deviations ≤ 0.02 for all cases

words, and non-German placenames, but ignored hybrid/bilingual compounds and English words with German morphology during evaluation. Our final system consists of a binary classifier augmented with dictionary lookup for known words and two routines to deal with German morphology (affixation and compounding).

The baseline classifier We used MALLET (McCallum, 2002) to train a maximum entropy classifier, using character 1- through 6-grams (including word boundaries) as features. Since we could not manually annotate a large portion of the MZEE corpus, the training data consisted of the disjoint subsets of the English and German CELEX wordlists (Baayen et al., 1995), as well as the words used in Covo (to obtain coverage of hip hop English). We tested the classifier using 10-fold cross validation on the training data and on a manually annotated development set of 10K consecutive tokens from MZEE. All data was lowercased (this improved performance). We excluded from both data sets 4,156 words shared by the CELEX wordlists (such as Greek/Latin loanwoards common to both languages and homographs such as *hat*), 100 common German and 50 common English stop words, all 3-character words without vowels and 1,019 hip hop artists/label names, which reduced the development set from 10K tokens, or 3,380 distinct types, to 4,651 tokens and 2,741 types.

Affix-stripping Since German is a moderately inflected language, anglicisms are often 'hidden' by German morphology: in *geflowt* 'flowed', the English stem *flow* takes German participial affixes. We therefore included a template-based affix-stripping preprocessing step, removing common German affixes before feature extraction. Because of the possibility of multiple prefixation or suffixation (e.g. *rum-ge-battle* ('battling around') or *deep-er-en* ('deeper')), we stripped sequences of two prefixes and/or three suffixes. Our list of affixes was built from commonly-affixed stems in the MZEE corpus and a German grammar (Fagan, 2009).

		Precision				
		All tokens		All types		OOVtyp.
Affix	Comp.	nodict	dict	nodict	dict	nodict
no	no	0.63	0.64	0.58	0.62	0.26
no	yes	**0.66**	0.69	0.58	0.62	0.27
yes	no	0.59	0.69	**0.60**	0.66	0.29
yes	yes	0.60	**0.70**	**0.60**	**0.67**	0.32

Table 1: Type- and token-based precision at recall=95

Compound-cutting Nominal and adjectival compounding is common in German, and loanword compounds are commonly found in MZEE:

(1) a. *chart*|*tauglich* ('suitable for the charts')
b. *flow*|*maschine*|*mässig* ('like a flow machine')
c. *Rap*|*vollpfosten* ('rap dumbasses')

Since these contain features that are highly indicative of German (e.g. *-lich#*, *ä*, and *pf*), we devised a compound-cutting procedure for words over length l (=7): if the word is initially classified as German, it is divided several ways according to the parameters n (=3), the number of cuts in each direction from the center, and m (=2), the minimum length of each part. Both halves are classified separately, and if the maximum anglicism classifier score out of all splits exceeds a target confidence c (=0.7), the original word is labeled a candidate anglicism. Parameter values were optimized on a subset of compounds from the development set.

Dictionary classification When applying the classifier to the MZEE corpus, words which occur exclusively in one of the German and English CELEX wordlists are automatically classified as such. This improved classifier results over tokens and types, as seen in Table 1 in the comparison of token and type precision for the dict/nodict conditions.

Evaluation We evaluated our system by adjusting the classifier threshold to obtain a recall level of 95% or higher on anglicism tokens in the development set (see Table 1). The final classifier achieved a per-token precision of 70% (per type: 67%) at 95% recall, a gain of 7% (9%) over the baseline.

Our system identified 1,415 anglicism candidate types with a corpus frequency of 100 or greater, out of which we identified 851 (57.5%) for further investigation; 441 (31.1%) were either established anglicisms, place names, artist names, and other loanwords, and 123 (8.7%) were German words.

4 Predicting the fate of anglicisms

We examine here factors hypothesized to play a role in the establishment (or decline) of anglicisms.

Frequency in the English Covo corpus We first examine whether a word's frequency in the English-speaking hip hop community influences whether it becomes more frequently used in the German hip hop community. We aligned four large (>1M words each) 12-month time windows of the Covo and MZEE corpora, spanning the period 11-2003 through 11-2007. We used the 851 most frequent anglicisms identified in our system to find 106 English stems commonly used in German anglicisms, and compute their relative frequency (aggregated over all word forms) in each Covo and MZEE time window. We then measure correlation coefficients r between the frequency of a stem in Covo at time T_t, $f_t^E(\mathit{stem})$, and the change in log frequency of the corresponding anglicisms in MZEE between T_t and a later time T_u, $\Delta \log_{10} f_{t:u}^G(w) = \log_{10} f_u^G(w) - \log_{10} f_t^G(w)$, as well as the corresponding p-values, and coefficients of determination R^2 (Table 2). There is a significant positive correlation between the variables, especially for change over a two-year time span.

Covo $\log_{10} f_t(\mathit{stem})$ vs. MZEE $\Delta \log_{10} f_{t:u}(\mathit{stem})$					
	r	p	t	R^2	N
$u = t + 1$ year	0.1891	*0.0007*	3.423	**3.6%**	318
$u = t + 2$ year	0.3130	*0.0001*	4.775	**9.8%**	212
$u = t + 3$ year	0.2327	*0.0164*	2.440	**5.4%**	106

Table 2: Correlations between stem frequency in Covo during year t and frequency change in MZEE between t and year $u = t + i$

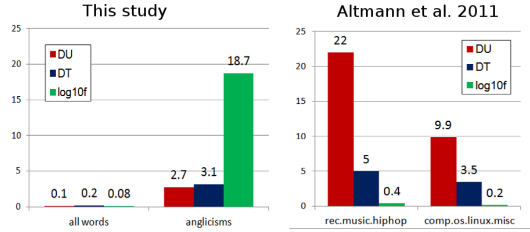

Figure 2: Correlation coefficient comparison of D_U, D_T, $\log_{10} f$ with $\Delta \log_{10} f$

Initial frequency and dissemination in MZEE In studying the fate of all words in two English Usenet corpora, Altmann, Pierrehumbert and Motter (2011, p.5) found that the measures D^U (dissemination over users) and D^T (dissemination over threads) predict changes in word frequency ($\Delta \log_{10} f$) better than initial word frequency ($\log_{10} f$). $D^U = \frac{U_w}{\tilde{U}_w}$ is defined as the ratio of the actual number of users of word w (U_w) over the expected number of users of w ($\tilde{U}_w$), and $D^T = \frac{T_w}{\tilde{T}_w}$ is calculated analogously fo the actual/expected number of threads in which w is used. $\tilde{U}_w$ and $\tilde{T}_w$ are estimated from a bag-of-words model approximating a Poisson process.

We apply Altmann et al.'s model to study the difference in word dynamics between anglicisms and native words. Since we are not able to lemmatize the entire MZEE corpus, this study uses the 851 most common anglicism word forms identified by our system, treating all word forms as distinct. We split the MZEE corpus into six non-overlapping windows of 2M words each (T_1 through T_6), calculate $D_t^U(w)$, $D_t^T(w)$ and $\log_{10} f_t(w)$ within each time window T_t. We again measure how well these variables predict the change in log frequency $\Delta \log_{10} f_{t:u}(w) = \log_{10} f_u(w) - \log_{10} f_t(w)$ between the initial time T_t and a later time T_u, with $u = t + 1, ..., t + 3$.

When measured over all words excluding anglicisms, $\log_{10} f_t$, D_t^U, and D_t^T at an initial time are very weakly ($0.0309 < r < 0.0692$), but significantly ($p < .0001$) positively correlated with $\Delta \log_{10} f_{t:u}$. However, in contrast to Altmann et al.'s findings that D^U and D^T serve better than frequency as predictors of word fate, for the set of anglicisms (Table 3), all correlations were both negative and stronger, and initial frequency $\log_{10} f_t$ (not dissemination) is the best predictor, especially as the time spans increase in length. That is, while most words' frequency change cannot generally be predicted from earlier frequency, we find that, for anglicisms, a high frequency is more likely to lead to a decline, and vice versa.[1].

[1] A set of 337 native German words frequency-matched to the most common 337 anglicisms in our data set patterns with the superset of all words (i.e., is not well predicted by any of the

$\Delta \log_{10} f_{t:t+1}(w)$					
	r	p	t	R^2	N
$\log_{10} f_t$	-0.2919	*<.0001*	-19.641	**8.5%**	4145
D_t^U	-0.0814	*.0001*	-5.258	**0.7%**	4145
D_t^T	-0.0877	*.0001*	-5.668	**0.8%**	4145
$\Delta \log_{10} f_{t:t+2}(w)$					
$\log_{10} f_t$	-0.3580	*<.0001*	-22.042	**12.8%**	3306
D_t^U	-0.1207	*.0001*	-6.987	**1.5%**	3306
D_t^T	-0.1373	*.0001*	-7.97	**1.9%**	3306
$\Delta \log_{10} f_{t:t+3}(w)$					
$\log_{10} f_t$	-0.4329	*<.0001*	-23.864	**18.7%**	2471
D_t^U	-0.1634	*.0001*	-8.229	**2.7%**	2471
D_t^T	-0.1755	*.0001*	-8.858	**3.1%**	2471

Table 3: Correlations between initial frequency and dissemination over users and threads and a change in frequency for the 851 most common anglicisms in MZEE.

Finally, from the comparison of timespans in Table 3, we see that the predictive ability (R^2) of the three measures increases as the timespan for $\Delta \log_{10} f$ becomes longer, i.e., frequency and dissemination effects on frequency change do not operate as strongly in immediate time scales.[2].

5 Conclusion

In this study, we examined factors hypothesized to influence the propagation of words through a community of speakers, focusing on anglicisms in a German hip hop discussion corpus. The first analysis presented here sheds light on the lexical dynamics between the English and German hip hop communities, demonstrating that English frequency correlates positively with change in a borrowed word's frequency in the German community–this result is not shocking, as the communities are exposed to shared inputs (e.g., hip hop lyrics), but the strength of this correlation is highest in a two-year timespan, suggesting a time lag from the frequency of hip hop terms in English to the effects on those terms in German. Future research here could profitably focus on this relationship, especially for terms whose success in the English and German hip hop communities is highly disparate. Investigation of those terms could suggest non-frequency factors which affect a word's success or failure.

The second analysis, which compared three measures used by Altmann, Pierrehumbert, and Motter (2011) to predict lexical frequency change, found that $\log_{10} f$, D^U, and D^T did not predict frequency change well for non-anglicism words in the MZEE corpus, but that $\log_{10} f$ in particular does predict frequency change for anglicisms, though this correlation is inverse; this finding relates to another analysis of loanwords. In a diachronic study of loanword frequencies in two French newspaper corpora, Chesley and Baayen (2010, p.1364-5) found that high initial frequency was "a bad omen for a borrowing" and found an interaction effect between frequency and dispersion (roughly equivalent to dissemination in the present study): "As dispersion and frequency increase, the number of occurrences at T2 decreases."

A view of language as a stylistic resource (Coupland, 2007) provides some explanation for these counter-intuitive findings: An anglicism which is used less often initially but survives is likely to increase in frequency as other speakers adopt it for 'cred' or in-group prestige. However, a highly frequent anglicism seems to become increasingly undesirable–after all, if everyone is using it, it loses its capacity to distinguish in-group members (consider, e.g., the widespread adoption of the term *bling* outside hip hop culture in the US). This circumstance is reflected by a drop in frequency as the word becomes passé. This view is supported by ethnographic interviews with members of the German hip hop community: *"Yeah, [the use of anglicisms is] naturally overdone, for the most part. It's targeted at these 15, 14-year-old kids, that think this is cool. The* crowd*! Ah, cool! Yeah, it's true–the* crowd*, even I say that, but not seriously."* -'Peter', 22, beatboxer and student at the Hip Hop Academy Hamburg.

In summary, the analyses discussed here leverage the opportunities provided by large-scale corpus analysis and by the uniquely language-focused nature of the hip hop community to investigate issues of sociohistorical linguistic concern: what sort of factors are at work in the process of linguistic change through contact, and more specifically, which word-extrinsic properties of stems and wordforms condition the success and failure of borrowed English words in the German hip hop community.

variables) in this regard.

[2] An analysis which truncated the forms in the first two timespans to match the N of the third confirm that this increase is not simply an effect of the number of cases considered.

Acknowledgements

Matt Garley was supported by the Cognitive Science/Artificial Intelligence Fellowship from the University of Illinois and a German Academic Exchange Service (DAAD) Graduate Research Grant. Julia Hockenmaier is supported by the National Science Foundation through CAREER award 1053856 and award 0803603. The authors would like to thank Dr. Marina Terkourafi of the University of Illinois at Urbana-Champaign Linguistics Department for her insights and contributions to this research project.

References

Beatrice Alex. 2008. *Automatic detection of English inclusions in mixed-lingual data with an application to parsing.* Ph.D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.

Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. 2011. Niche as a determinant of word fate in online groups. *PLoS ONE*, 6(5):e19009, 05.

R.H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. The CELEX lexical database. CD-ROM.

Paula Chesley and R.H. Baayen. 2010. Predicting new words from newer words: Lexical borrowings in french. *Linguistics*, 45(4):1343–1374.

Nikolas Coupland. 2007. *Style: Language variation and identity*. Cambridge, UK: Cambridge University Press.

Sarah M.B. Fagan. 2009. *German: A linguistic introduction*. Cambridge, UK: Cambridge University Press.

Hans Henrich Hock and Brian D. Joseph. 1996. *Language history, language change, and language relationship: An introduction to historical and comparative linguistics*. Berlin, New York: Mouton de Gruyter.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Web: http://mallet.cs.umass.edu.

Alexander Onysko. 2007. *Anglicisms in German: Borrowing, lexical productivity, and written codeswitching*. Berlin: Walter de Gruyter.

Alastair Pennycook. 2007. *Global Englishes and transcultural flows*. New York, London: Routledge.