CROSSING THE LEXICON: ANGLICISMS IN THE GERMAN HIP HOP COMMUNITY

BY

MATTHEW E. GARLEY

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Linguistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Associate Professor Marina Terkourafi, Chair
Assistant Professor Julia Hockenmaier
Professor Rakesh Bhatt
Professor Hans Henrich Hock
Professor Jannis Androutsopoulos, Universität Hamburg
The influence of English on German has been an ongoing subject of intense popular and academic interest in the German sphere. In order to better understand this language contact situation, this research project investigates anglicisms—instances of English language material in a German language context—in the German hip hop community, where the use of novel anglicisms is especially frequent. This investigation takes a methodologically diverse approach, including complementary corpus, sociolinguistic, and ethnographic analyses. In this dissertation, I focus primarily on an original 12.5-million-word German-language corpus of hip hop discussions from the Internet forums at MZEE.com which includes 11 years of computer-mediated discourse. I supplement these data with an English-language hip hop discussion corpus and a set of ethnographic interviews conducted with hip hop fans and artists in Hamburg in the summer of 2010.

I first detail the development of a computational classifier which identifies novel anglicisms in the MZEE.com corpus with high accuracy, yielding a list of 850 frequent anglicisms which is in turn used to identify unexpected wordforms—those which have a non-canonical morphological or orthographic nativization. Through an exploration of the linguistic properties, frequency, and distribution of these forms, I demonstrate the close link between orthographic, morphological, and phonological expressions of these anglicisms and argue that these forms are the result of extraordinary interaction of German and English linguistic-orthographic rules. The next analysis investigates the diachronic fate of anglicisms in the MZEE corpus, finding that frequency in an initial time window is significantly, and negatively, correlated with change in frequency for the set of 850 anglicisms—and this correlation is much stronger for anglicisms than for native German words, indicating the limited shelf life of anglicisms' stylistic utility, a situation corroborated by the subsequent analysis of ethnographic interviews with linguistic actors in the German hip hop community. That analysis reveals systematic and enduring constellations of attitudes toward German, English, and the use of anglicisms which interact with what I term the standard language ideology complex for German (including the related ideologies of the standard language, language purism, and Herderian ideology)—finding a surprising basis of linguistic conservatism which, even when opposed by individual actors, seems to reliably frame metalinguistic discourses.

In combination, these findings, 1) that the nativization of anglicism wordforms is rule-governed, even when it appears haphazard or disruptive; 2) that many novel anglicisms seem to have a limited timeframe of popularity; and 3) that the standard language ideology complex and other related ideological stances toward anglicisms are dominant, even in a subcultural community where English material is ubiquitous and linguistic 'resistance' is hypothesized; suggest that concerns about the imminent decline or loss of the German language are (to borrow from Mark Twain, an infamous student of the language) 'an exaggeration'.
To my mom, my dad, and my grandparents, who always knew I'd make it.
I wasn't always certain I'd finish this dissertation. That I did is thanks to those mentioned here, whose guidance, encouragement, support and friendship I can only hope to repay.

My advisor, Marina Terkourafi, took a chance on me as a student and patiently let me forge my own path. Marina has been a wonderful and committed mentor, and in fact organized, in response to student interests, the workshop on language and hip hop where my very first research on the subject was presented. Since that time, Marina has been not only an insightful and painstaking reviewer of my work, but a top-notch advisor and supporter, directing me to productive avenues of research, presentation, and publication.

I also owe a great debt of thanks to Julia Hockenmaier, who welcomed a student with limited computational knowledge and fostered the development of those skills into a much more robust understanding of the computational aspects of my project. Over the course of several years, I have greatly appreciated the opportunity to collaborate and exchange ideas with Julia, and her patience with my developing abilities and her willingness to expend great sums of time and effort on my behalf were crucial in my ability to bring this project to completion.

Rakesh Bhatt has been a friendly face in the Linguistics department since my first day of graduate study, and has since served as a knowledgeable and engaging expert and at the same time a wonderful and encouraging teacher in sociolinguistics, and I attribute my keen interest in language contact to experiences in several of his courses.

Hans Henrich Hock has also played a crucial role in the development of this thesis, contributing a deep understanding of linguistic change and an unparalleled knowledge of the relevant literature. I have been honored to work with such an esteemed scholar, and have benefited greatly from his suggestions and comments on my work.

Finally, I am grateful to Jannis Androutsopoulos for his generous invitation to supervise my research project in Hamburg. I have had productive and enlightening interactions with Jannis, who is among his many other accomplishments, an inspiring author at the intersection of hip hop and computer-mediated communication.

There are a number of other academics who have made great contributions to this dissertation by adding to my understanding of certain subjects, by asking thought-provoking questions, or simply by providing encouragement along the way. I would like to begin by thanking Susan Herring, who has provided a voice of encouragement at various conferences and who I feel has included me in the research community. Others who have made their mark on this thesis by reading my work, providing conference feedback, or simply commenting on various aspects of my work include, Michele Koven, Richard Sproat (now at OHSU), Janet Pierrehumbert at Northwestern, Alexander Onysko at Innsbruck, and Ercan Balcı. I would also like to thank a number of faculty at Illinois and elsewhere who have given me encouragement, advice, and support along the way: Andrea Golato, Karlos Arregi, Jennifer Cole, Abbas Benjemoun, Peter Lasersohn, Dennis Baron, James Yoon, Eyamba Bokomba, Jonathan MacDonald, Randall Sadler, Rick Hallett at NEIU, Lars Hinrichs at UT Austin, Lauren Squires at The Ohio State University,
I am deeply grateful for the support of other graduate students in (and around) my field at Illinois, a great number of whom are now faculty and professionals in their own right. Many made contributions to this dissertation, and all have provided the kind of positive diversion that refreshes academic creativity: Benjamin Slade, collaborator (in the good sense) in many ways and boon companion; Liam Moran, for writing the script to collect the original corpus; Karen Lichtman, for helping me (finally) figure out how to structure my time; Jill Hallett; Nikos Vergis; Casey and Stefanie Coughlen; Lisa Pierce; Daniel Scarpace; Ern Rusaw; Soondo Baek; Eunah Kim; Tim Mahrt; Amanda Huensch; Lori Moon; Jennifer Cramer; Antje Muntendam; Erica Britt; Adriana Molina-Muñoz; Sarah Simeziane; Aimee Alnet; Gary Linebaugh; Tae-jin Yoon; Vandana Puri Sharma; Kevin Stillwell; Andrew Fister; Andrew Hinderliter; Marco Shappeck; Eman Saadah; Chris Stewart; and Sarah Henneböh for proofreading quite a bit of German transcription. I am also very grateful to a number of friends outside the field: I would like to thank Boris Pilev, Stanimir Kondov, Iva Yankova, and Radoslav Kirov (Həzərəwə!) Nick Brönn, Dave Wells, Colleen Cook, Josh Lüdtke, David Holmes, Ellen Smith, Dave, Jared, Steve, Jeff, and the rest of the Armored Gopher crew (for therapeutic dice-rolling and goblin-smashing), Matt Newton, Brandon Miller, Josh Stowell, Adam Noonchester, Eric Noonchester, Annie Keatley (Horch, Urfriend!), and Tess Halonen (for brilliantly explaining my statistics to me when I didn't even understand them).

My research in Germany would not have been possible without the friendly and welcoming contacts inside and outside of the hip hop community: Christoph Baumeister, my first and best contact throughout this project; Johannes Jacobasch and the students of the Hip Hop Academy Hamburg, DJ Mirko Machine, Beat Boy Delles, and Sleepwalker; the artists and hip hop fans in Hamburg who agreed to impertinent interviews about language conducted by an American, and finally Benedikt Mäder, Meiken Ehlers and their friends, for hospitality and generosity of spirit.

Without the excellent support staff in Linguistics and the School of Literatures, Cultures, and Linguistics, my progress in the program would not have been possible, and I owe them a great deal: Mary Ellen Fryer, Pat Gallagher, Geraldine Moore, and Cathy Penny.

Throughout my graduate study and research, I have been supported by a number of organizations, to whom I am grateful: the Graduate College and the Department of Linguistics at Illinois through a number of fellowships and assistantships, the German Academic Exchange Service (DAAD) through a Graduate Research Grant, and the Beckman Institute at the University of Illinois through a Cognitive Science/Artificial Intelligence award.

Finally, I would like to thank my family, who have always been there for me in many ways.

I am sure to have overlooked some who were crucial in this endeavor—to those not named here who have played a positive role, direct or indirect, in my work; you have my gratitude.
# Table of Contents

CHAPTER 1: INTRODUCTION ........................................................................................................... 1

CHAPTER 2: THEORETICAL PRELIMINARIES: A CLASSIFICATORY SCHEME FOR ANGLICISMS IN GERMAN .......................................................................................................................... 12

CHAPTER 3: THE CORPUS DATA .................................................................................................... 24

CHAPTER 4: THE AUTOMATIC DETECTION OF ANGLICISMS IN LARGE CORPORA ................................................................................................................................. 37

CHAPTER 5: LOANWORD ADAPTATION IN THE GERMAN HIP HOP COMMUNITY ................................................................. 70

CHAPTER 6: PREDICTING CHANGES IN ANGlicism USE ........................................................................ 102

CHAPTER 7: THE ROLE OF IDEOLOGY IN LINGUISTIC PRODUCTION AND STYLIZATION .................................................................................................................. 125

CHAPTER 8: CONCLUSIONS ............................................................................................................. 171

REFERENCES ................................................................................................................................ 184

APPENDIX A: LIST OF HIGH-FREQUENCY ANGLICISMS IN THE MZEE CORPUS .......................................................................................................................... 190

APPENDIX B: EXCERPTS FROM A FORUM DISCUSSION OF AN ORTHOGRAPHIC WORDFORM CONTAINING <ED> .............................................................................................................. 197

APPENDIX C: FINAL INTERVIEW PROTOCOL .................................................................................. 200

APPENDIX D: ORAL CONSENT PROTOCOL SCRIPT ........................................................................ 203

APPENDIX E: INSTITUTIONAL RESEARCH BOARD APPROVAL LETTERS .................................. 204

APPENDIX F: TRACKLIST/DISCOGRAPHY FOR EPIGRAMS BY CHAPTER ................................. 207
CHAPTER 1
INTRODUCTION

Ich bin halt für mehr Inhalt, denn ich vermiss' es
Ich hoff', dass jedes Wörtchen Hip Hop ein Begriff ist
Was? Ich soll mich verkaufen? Vergiss es, denn
so verwandelt sich Kultur zum Business

I stand for more content, because I miss it
I hope that every word of hip hop is an idea
What? I should sell myself? Forget it, because
that's how culture changes into business

–Umse, 'Kultur', 2008
(with lines sampled from Ruhrpott AG
and Kinderzimmer Productions)

1.1 The research problem: English in Germany

Language change has long been one of the most popularly misunderstood aspects of human
language; those who take language change as evidence of linguistic or cultural decline count
many highly-educated individuals among their number. In this chapter, I discuss these kinds of
perceptions of English influence on the German language, thereby framing the primary research
problem addressed in this dissertation and stating the overarching research questions. In the next
section, I will introduce and motivate the selection of the research domain, language use in the
German hip hop community. Finally, in the concluding section of this chapter, I will provide an
overview of the remainder of this dissertation, elucidating the connection between the key
questions and the studies included here that these questions motivate.

Controversy surrounding language change is evident in the historical records of both the U.S.
and Germany, where language debates between strict prescriptivists, die-hard descriptivists, and
linguistic commentators along the spectrum between these two extremes have raged in
dictionaries, usage guides, newspaper columns, and more recently, websites and blogs. Modern
linguists have, by and large, reached a consensus on the matter; Aitchison (1991: 221) writes that
“[language change] is natural and inevitable, and is due to a combination of psycholinguistic and
sociolinguistic factors”, further finding “[…] no evidence that language is either progressing or
decaying.”
The subject grows more complex when another language is brought into the fray. Contact-induced change—language change attributable to contact between languages—not only inspires fear of language decay or decline,¹ but often involves ideological rhetoric, like that of Drews (1999: 15), who describes extensive English borrowing in German as:

\[ \text{einer Invasion, einem Hereingeschwemmtwerden, einer massenhaften Übernahme englischer bzw. Angloamerikanischer Wörter, Formulierungen und Redewendungen.} \]

an invasion, a sucking-in, a mass adoption of English and Anglo-American words, phrases, and expressions. [my translation]

The word which crucially captures the concern in this statement is Invasion, which calls the reader to action in defense of the mother tongue. The ongoing debates about language change in the U.S. and elsewhere demonstrate that while linguists may be largely of one mind about the consequences of language change, other academics like Drews, a renowned literary critic and university professor, have remained unconvinced.

While the consequences of language change are fairly well understood within the discipline of linguistics, a full explanation of the various processes involved is lacking—and a more thorough understanding of these processes can provide insight into what, if anything, linguistic conservatives have to fear. Take the case of borrowing, one form of contact-induced change: the forces driving borrowing at a general level are known—Hock & Joseph (1996: 271) present a concise overview, identifying need (in the case of concepts which don't have corresponding linguistic expressions in the recipient language) and prestige (i.e., exploitation of another language's social cachet) as primary motivations for borrowing. However, the specific mechanisms of borrowing on the individual and community level remain largely unexplained, and the study of these mechanisms, including the role of the individual in the borrowing process, the factors which influence the dissemination of borrowed material through a community, and

---

¹ It is important to draw a distinction between the phenomenon called 'language decay' by some linguists (a shift from the use of one language to another, and eventual loss or death of the original language) and qualitative evaluations of language change by prescriptivists, the latter of which I refer to here. The former, however, also plays a role in popular fears about language change. For a discussion of the distinction between the (stable) situation of contact-induced change and the phenomenon of language decay (nigh-inconceivable in the German case), see Sasse (1992).
the social circumstances which foster or hinder borrowing, promises to shed light not only on popular dialogues about language change but on enduring questions of sociohistorical linguistics.

A researcher in the social sciences serves two communities. The first, the academic community, is served by the researcher's advancement of knowledge on a particular set of subjects, which should inform and challenge further research. The second, the public, is the ultimate beneficiary of social science research, and it is a commonly expressed sentiment that academic researchers would do well not to answer unasked questions. In this spirit, I present here an excerpt from the volume *Sprache in Not? Zur Lage des heutigen Deutsch* 'Language in Danger? On the status of modern German'. After providing a litany of examples of purported English influence on German, Meier (1999: 8) writes:


What does this all mean? What is happening in and with our language? “The power of a language is not in rejecting the foreign, but rather in devouring it”, found Goethe (Maximen und Reflexionen Nr. 736). But devouring is still a form of appropriation, whereas today, one could ask if we haven't already had at least something to do with a divestiture of our language. [my translation]

Meier, along with Drews (1999) and others, raise the question: does English influence pose a threat to the German language? This dissertation sheds light on this question, the relevance of which has not waned since the publication of Meier's volume in 1999. In August of 2009, the CDU, the political party of German Chancellor Angela Merkel, sponsored a discussion at Frankfurt's International Theater entitled *Sorge um die deutsche Sprache*, 'Concern about the German language', with two questions posed in the event's description: “Do anglicisms portend the fall of the German language?” and, “Will identity be forfeit?” (Frankfurter Rundschau 2009).

---

2 Throughout this dissertation, I adhere to the following typographic conventions:
   a) Italic typeface is used to indicate terms introduced or used in a technical sense, as well as foreign terms and all words used in a German-language context (including borrowed English words). English translations, when appropriate, are presented immediately following in single quotes, except when presented as block citations.
   b) Underlining is used in the main text for emphasis and for the titles of published volumes.
   c) Bold face is used to call attention to words from longer examples which are of special relevance to the immediate discussion (in general, English-derived words in a German context).
The connection between language and identity that this second question assumes is central to the discipline of sociolinguistics, and serves as a jumping-off point for the research agenda of this dissertation. In the service of answering questions about the threat that English influence might pose to German, and in the interest of furthering scientific understanding of the interaction of culture and linguistic change, I am concerned with four overarching questions related to the central theme of contact-induced change. First, can anglicisms—using the working definition 'English linguistic material in a German context' (for further discussion, see Chapter 2)—be reliably identified by automatic means? Second, what, if anything, constrains the borrowing process with regard to the selection of language material for borrowing and the nativization or adaptation of borrowed forms, or to put it differently: what gets borrowed, and in what form? Third, what causes anglicisms, and especially lexical borrowings, to succeed or fail over time? Finally, what are the ideological motivations and social connections that condition the use of anglicisms?

This is not the first time these questions have been posed: The first can be considered a specialized case of the computational application of language identification—see, e.g., Newman (1987) and Beesley (1988). However, the classic language identification problem focuses on the identification of the primary language in an entire document; only more recent approaches like Alex’s (2008) English inclusion classifier have engaged the problem at the lexical level. The second question, which regards the types of language material which can be borrowed and the degree and manner of nativization (addition of inflectional morphology, phonological change, etc.), has long been a concern of those studying language: In 1854, the famous German philologist Jacob Grimm, noted:

*Fällt von ungefähr ein fremdes wort in den brunnen einer sprache, so wird es solange darin umgetrieben, bis es ihre farbe annimt und seiner fremden art zum trotz wie ein heimisches aussieht.*

When a foreign word falls by accident into the fountain of a language, it will get driven around in there until it takes on that language's colour and resembles a native term in spite of its foreign nature.

[cited in translation by Onysko (2007: 60)]

Recent work on what can be borrowed and the nativization of borrowed words includes Muysken's (2000) and Field's (2002) comprehensive treatments of borrowing. I reserve detailed
discussion of these works for Chapter 3. With regard to the third question, the dynamics of lexical items over time in a large corpus was the focus of the study performed by Altmann et al. (2011), but this study did not focus on borrowings; Chesley & Baayen's (2010) analysis did focus on borrowings (in French) but used a smaller newspaper corpus. Both studies found evidence that factors like relative frequency and the distribution of lexical item use over speakers or portions of text could, to a fairly small extent, predict the later success or failure of a lexical item. With regard to the final question, the investigation of the social meanings and ideological motivations associated with linguistic forms has been carried out in a number of classic and influential studies. One widely-cited example is Gal's (1979) Oberwart study, which is especially notable for its ethnographic approach that allowed for the collection of naturalistic data linking situational code choice (German or Hungarian) to complex social factors involving changes in the community, which in turn contributed to a comprehensive analysis describing and explaining interlocking social and linguistic processes on the large and small scale.

Neither are these long-settled questions. Active research continues in the areas outlined by these questions: Johnstone et al. (2006) discuss social factors and the development of a dialect in Pittsburgh along Silverstein's (2003) orders of indexicality, and (national) identity's role in dialect formation has even been called into question recently, as Trudgill (2008) and the round of responses in the same volume attest. In terms of the spread of linguistic forms through a community, Milroy & Milroy's (1985) line of research is continued in Paolillo's (2001) analysis of multilingual social networks in an Internet chatroom. A related approach is exemplified in Eckert's (2000) analysis of stylization at a Detroit high school, which crucially involves participation in an ongoing sound change. This approach takes a community to be constituted by a shared repertoire of practices, which, like social network-based analyses, considers language use on both the individual and community level.

Over the last century and a half, then, linguists have carried out research with these kinds of questions in mind; however, this fact does not make the questions any less relevant today. Public interest in language change and its consequences far outstrips public knowledge about language change. From the standpoint of the discipline, the confluence of new data sources, new theories and methodologies of sociohistorical linguistics—as well as new populations and domains for
study—makes it likely that far more nuanced, and possibly entirely different, answers are to be found. The promise that different sources of data might shed light on these enduring questions in part motivates the choice of material for study, which I address briefly in the next section and in detail in Chapter 3.

1.2 The German hip hop community as a research domain

For the present research, I have chosen the German-language hip hop community as the site of investigation, focusing primarily on a large corpus of Internet forum discussions and a smaller collection of in-person interviews and correspondence with fans and hip hop artists. Overall, the majority of the data in both sources come from hip hop fans. This choice of data reflects a convergence of multiple related factors. First, German-language data were chosen due to the prominence of recent public debates regarding perceptions of English influence on the German language, a kind of concern about language which I re-iterate here is not entirely unique to Germany, but which has a very long historical, political, and cultural context in the German-speaking sphere (Spitzmüller 2007). Second, an analysis of language in the German hip hop movement relates to ongoing research on the emergence of global hip hop, which has been a recent subject of interest in sociolinguistics and related fields (see, e.g., recent volumes from Pennycook (2007), Alim et al. (2009), and Terkourafi (2010)). In addition, Alim (2004: 388) has noted the centrality of linguistic behavior to the practice of hip hop (in the form of rapping, one of hip hop's traditional 'four pillars'), a placement making hip hop an ideal ground for the study of linguistic transfer. However, linguistic analysis of rap lyrics poses a conundrum to the researcher: rap lyrics are often carefully composed, edited, and rehearsed; and while extemporaneous compositions exist (a type of rapping known as 'freestyling'), the demands of meter, rhyme, and other concerns of a highly codified musical genre distance rap lyrics from more prototypical forms of natural language production. In selecting data for analysis, then, I focus instead on everyday language use among German hip hop fans and artists. I select this type of data for several key reasons. First, it is important to consider remarks like those by Bentahila & Davies (2002: 193) in their discussion of codeswitching, a language contact phenomenon, in the lyrics of North African rai music:
The self-conscious exploitation of switching by song writers can no more be taken as
evidence for how ordinary speakers unconsciously switch in conversation than, say, the
poetry of e. e. cummings can be used as a reference for studying English punctuation.

In particular, because of the centrality of meter and rhyme scheme to the practice of rapping, it is
not inconceivable that the use of an English or German word in a given context has less to do
with the source language than with the completion of a rhyme or the proper number of syllables
for a meter.

Androutsopoulos (2010) builds on this point, noting specifically that rap lyrics go through
successive stages of editing and are, in addition, public (rather than intimate) discourse. Sarkar
(2009), on the other hand, puts forth a different view, arguing that rap lyrics can faithfully reflect
the speech of the hip hop community, and that if multilingualism in rap lyrics is not understood
by the (relatively small) target audience, that audience will adapt by gaining the requisite
knowledge. Sarkar (p.c.), however, cautions that she believes this claim to be applicable in the
case of her own data, and that the legitimacy of such a claim must be established separately for
any given sociolinguistic situation. She points out that it is at least true that hip hop artists claim
to rap like they speak, and (presumably) believe the same. This is a valid argument, but given the
choice between more naturally-occurring and spontaneous linguistic production and less
naturally-occurring and spontaneous linguistic production, it seems most reasonable to study the
former. While investigation of the latter is indeed warranted, lyrics are an area which has been
quite commonly studied with regard to global hip hop—leaving out parallel, and no less
important, types of discourse which properly belong to hip hop; this is the second reason for
focusing on informal linguistic productions rather than on lyrics. Noting a number of
understudied subdomains in hip hop, Androutsopoulos (2009: 44) calls for the exploration and
integration of these largely ignored territories in linguistic research on hip hop:

[...] most language-centered studies on Hip Hop focus on rap lyrics. Although this focus has yielded many important results so far, it seems to overlook the emic distinction between Hip Hop as a cultural hyperonym and rap as one of its hyponyms [...] An integrative view on language and Hip Hop would need to encompass a much wider range of discourse practices, such as talk at work among rappers, writers, and breakers; the discourse of Hip Hop magazines and broadcast shows; artist-fan communication during live events; and an array of everyday talk and computer-mediated discourse in what is often termed the Hip Hop Nation.

[emphasis mine]
Finally, in attempting to answer questions about how language change functions, it is crucial to examine fans' and artists' everyday usage in order to gauge whether the primary texts, i.e., English or English-influenced rap lyrics and other mediated, globalized discourses, are taken up to any significant extent within the hip hop community.

To summarize, there are multiple key motivations for my choice of data: the prominence of linguistic debates regarding anglicisms in German, the unique nature of global hip hop language, the (disputed) lack of equivalence between lyrics and everyday language production, the understudied nature of non-lyrical aspects of global hip hop, and the necessity of looking at everyday language use in order to find English influence on German.

In this dissertation, I engage anglicism use in the German hip hop community through a multifaceted research project drawing on multiple types and sources of data, triangulating an accurate, comprehensive, and situated view of linguistic practices. The largest portion of data is comprised of natural language corpora collected from a 12.5 million word German-language Internet hip hop discussion forum covering an 11-year period and a 19.5 million word American (English-language) hip hop discussion forum. Corpora of this type lend themselves especially well to large-scale quantitative analyses of borrowing patterns and smaller-scale linguistic analysis, but the analysis of corpus data alone has a major drawback in that the key social and ideological motivations underlying the patterns found in such data must remain, at best, speculative. For this reason, I combine linguistic and corpus-linguistic analyses with ethnographic, in-person interviews with members of the hip hop community, in which expressed attitudes about language reveal the ideologies and motivations underlying and conditioning the use of anglicisms. By integrating these analyses, this dissertation provides a comprehensive investigation of the research object: anglicisms in the German hip hop community. The investigation of anglicisms in turn provides a window into the specific ways in which contact-induced language change operates and illuminates the initial research problem—the extent to which English might be said to be influencing, changing, or existentially threatening, German. In the next and final section of this chapter, I outline the structure of the remainder of this dissertation and discuss the research questions addressed in each of the studies undertaken here.
1.3 Research questions and overview of chapters

In order to properly engage in the study of anglicisms, a logical first step is the clear definition of this research object. I engage with this question in Chapter 2 by reviewing crucial background material regarding language contact phenomena and developing a robust understanding of what is meant by the notion of anglicism. Following this, in Chapter 3, I will introduce Internet forums, from which a large portion of the data derive, as a type of computer-mediated discourse. In addition, I will discuss the preparation of the corpora used here, provide thorough descriptions of the particular corpora collected for this dissertation, and report on a preliminary by-hand analysis which establishes the relative frequency and kind of anglicisms found in the primary corpus.

The ultimate catalyst for the research agenda of this dissertation, as mentioned in the previous section, is the question—posed in public linguistic debates—of whether influence from English constitutes a threat to the German language. I therefore seek to determine whether and to what extent this is the case, and perhaps more realistically, what might lead to the perception of this threat. What is clear is that German is being influenced by English in some way; I also seek to more precisely determine the ways in which this influence is proceeding. In order to address these overarching concerns, I pose the following research questions, first introduced in section 1.1:

1. Can anglicisms be reliably identified by automatic means in natural, highly contextualized data (like the linguistic production in online forums)?

2. How does the nativization of anglicisms proceed, and what factors condition this process?

3. What factors cause anglicisms to fail or succeed over time, and what does this tell us about their functions?

4. What are the underlying and motivating ideologies that affect the use (and possibly the adaptation) of anglicisms?

In order to be able to address these issues using large-scale corpus data, I begin by addressing the first of these questions through the development of an anglicism classifier in Chapter 4. The automatic identification of anglicisms in a large corpus opens the door for analyses of anglicism
properties and patterns of use. However, the ramifications of developing such a system go beyond the applied; the affordances that must be built into such a system in order to reliably capture nativized forms (for example, stripping certain affixes off of an English stem or dividing compound wordforms in order to classify their components) already begin to illuminate the way in which anglicisms are generally adapted into German. In the end, the classifier succeeds in providing a list of candidate anglicisms with high recall (picking out over 95% of the anglicisms identified by human annotators in a development set) and good accuracy. By narrowing the wide range of wordforms in the large corpus to a much smaller list which it was feasible to hand-correct, the classifier yields a set of anglicisms that occur 100 times in the corpus (a frequency of roughly 8 occurrences per million words).

The reliable automatic identification of a large number of anglicism forms in the corpus allows for engagement with the second research question in Chapter 5, where, after discussing the generally accepted and previously described ways in which nativization occurs, I present an analysis of the linguistic and orthographic properties of anglicisms taking unexpected or non-straightforward nativized forms. This analysis demonstrates that anglicisms' morphological, orthographic, and phonological nativization may at times appear haphazard or illogical, but that these aberrant forms are, in fact, conditioned by rules from one or both languages, which interact in unexpected ways. In addition, the engagement with the question of nativization underscores the utility to linguistic researchers of attention to orthography, which can sometimes reflect otherwise imperceptible linguistic processes.

In Chapter 6, I address the factors which play a role in the frequency of anglicism use by comparing the lexical dynamics of the identified anglicisms with those of other words in the German corpus, as well as by comparing the lexical dynamics of anglicism stems in the German hip hop corpus with those in an English hip hop corpus. The primary thrust of this study involves the prediction of anglicisms' fate (as gauged by the wordforms' relative increase or decrease in frequency over time), an investigation that underlines the ways in which anglicisms pattern differently from German words in the corpus, particularly highlighting the tendency for frequent anglicisms to decline in frequency and for infrequent anglicisms to rise in frequency over time.
The fourth and final research question is the primary focus of Chapter 7, which addresses speaker attitudes toward anglicisms and the language ideological processes which impact anglicism use. The analysis draws on ethnographic interviews conducted among hip hop fans and artists in Hamburg, Germany, as well as on selected Internet forum discussions on the topic of anglicisms, revealing the nature of the relationship between the standard language ideological complex and alternative ideologies in the hip hop community. This inquiry provides crucial explanations for linguistic and distributional facts presented in Chapters 5 and 6. The correspondence between anglicism distribution, the properties of anglicisms, and the explanations for these facts provided by the circulation of language ideologies in the hip hop community are the focus of Chapter 8, the concluding chapter. Chapter 8 also addresses the implications of the studies I undertake in this dissertation for theories of language contact and change.
CHAPTER 2
THEORETICAL PRELIMINARIES: A CLASSIFICATORY SCHEME FOR ANGLICISMS IN GERMAN

Es ist der big bad Boss, get back, Dogg
Smoke dein Sticky weg, zip dein Six-pack off, Homie
Ich geb kein Fuc, spit mein Stuff
und push mehr Heavyweightshit wie im Fitnessclub

It's the big bad boss, get back, dogg
Smoke your sticky up, zip your six pack off, homie
I don't give a fuck, spit my stuff
and push more heavyweight shit than in a fitness club

–Franky Kubrick, '110-112', 2004

2.1 Overview

In Chapter 1, I identified the research object of this dissertation as the use of anglicisms in the German hip hop community, using a working definition of anglicisms as 'English material used in a German-language context'. In this chapter, I examine previous treatments and systems of classification proposed in the literature for such language-contact phenomena, motivating my continued use of the term 'anglicisms' and more precisely specifying what is intended by its use. One of the key problems here is that some of the wordforms I discuss in this dissertation as anglicisms might not fit an intuitive definition of linguistic borrowing and might instead more commonly be seen as codeswitches or components thereof. In section 2.2, I establish a firm theoretical definition of linguistic borrowing which fits the largest number of English inclusions that I refer to as anglicisms. In section 2.3, I provide a discussion of borrowing with respect to the concepts of codeswitching and codemixing and whether it is possible or desirable to distinguish these language contact phenomena from the process of borrowing. Finally, in section 2.4 I present a classificatory schema for language contact phenomena and clarify the criteria used to determine whether a linguistic entity is to be considered an anglicism for the purposes of this research.

2.2 Borrowing as a language contact phenomenon

Haugen's (1950) analysis of borrowing is generally considered the first modern treatment of the subject. Haugen (1950: 211-212) establishes a terminology for the study of borrowing, advocating its use over terms suggesting linguistic 'mixture':
The real advantage of the term 'borrowing' is the fact that it is not applied to language by laymen. It has therefore remained comparatively unambiguous in linguistic discussion, and no apter term has yet been invented.

Unfortunately, the use of the term 'borrowing' in the years following Haugen's suggestion has been at times anything but unambiguous. Returning to Haugen's main thrust, we find his definition—which, he notes, has its own roots in Paul's (1886) discussion of borrowing—as follows: ‘Borrowing is then the attempted reproduction in one language of patterns previously found in another’ (Haugen 1950: 212). This scenario tacitly presuppose a speaker who has learned patterns from two languages and is thus, in this broad sense, bilingual. Bilingualism, then, is a prerequisite for borrowing, which Haugen further considers in its strictest sense to be a process carried out by individual speakers. Haugen contrasts borrowing as a process with loans, which are, per Haugen, the products of borrowing. In this framework we thus find one instance of a distinction made more generally in the literature on language change between diffusion and transmission (Labov 2007). Diffusion, on the one hand, can be conceived of as the reproduction of patterns across the boundaries of linguistic varieties. Transmission, on the other hand, involves the subsequent spread of a pattern; Labov (2007: 345-346) characterizes it as a passing-on of linguistic forms within a speech community. This distinction is key to understanding the difference between the process of borrowing, which requires some bilingual competence on the part of the borrower, and the use of a borrowed form, which does not require knowledge of a second language variety.

Definitions of borrowing in the more recent literature are provided by numerous authors, but many, including McMahon (1994: 200) and King (2000: 82), adopt Haugen's definition unproblematically. A number of others provide definitions which are quite compatible with Haugen's, if put differently (and given the texts' varying aims, this is to be expected). Hock (1991: 380) sees borrowing as “the adoption of individual words or even of large sets of vocabulary items from another language or dialect.” Fischer (2008: 6) writes:

\[ \text{Borrowing denotes the process as well as the object. As a process it typically refers to the importation of a word or its meaning from one language into another. As an object, it denotes the form and/or the meaning of the item that originally was not part of the vocabulary of the recipient language [...]} \]

[emphasis Fischer’s]
Onysko (2007: 79) assigns the label of borrowing to a subset of cases where both form and meaning are transferred from a source language (SL) to a receptor language (RL). While Onysko is in the minority in his use of this more restricted definition, the distinction between different types of language transmission (in other literature, different types of borrowing) on the basis of form and meaning is traditional. Haugen (1950) uses form and meaning, as well as the possibility of one being borrowed without the other, to classify loans into three types based primarily on the amount of form preserved in the process of borrowing: loanwords, which involve importation of both form and meaning, loanblends, that is, 'hybrids' which meld morphological forms from two languages into a single word, and loanshifts, which involve no transfer of form but only transfer of meaning (Haugen's loanshifts are more commonly referred to as semantic borrowings or calques today).

2.3 Classificatory schemata for language contact phenomena

Other authors have categorized language contact phenomena in different ways, often finding it necessary to define codemixing/codeswitching as complementary to borrowing. In these cases, two primary heuristics for distinguishing between these concepts emerge from the literature on the subject: the frequency of use, and the unit size of the language material used (in terms of either length in words or syntactic status of unit as a word/phrase/clause). Muysken (2000) notes in his review of the relevant literature that the models proposed by researchers studying these language contact phenomena are often influenced by the specific language pairs and situations which they study. Similarly, a distinction can be drawn between researchers who primarily engage the phenomenon of codeswitching and deal with borrowing as a special case (e.g., Myers-Scotton 1992) and those who concentrate primarily on the definition and study of linguistic borrowing, mentioning codeswitching as a secondary concern (e.g., Poplack et al. 1988; Onysko 2007).

The earliest researchers to operationalize frequency of use as a criterion for distinguishing between (what would intuitively be understood as) code-switching and borrowing are Poplack et al. (1988), who drew a distinction between established borrowings and nonce borrowings, with the latter defined as borrowings which appeared only once in a large (roughly 2 million word) corpus of spoken language. The most important finding from this study is that the use of
established borrowings vs. nonce borrowings varies from individual to individual along several social dimensions, and that a parallel between nonce borrowings and codeswitching can be drawn which distinguishes both of these from established borrowings, according to synchronic use. However, as discussed below, Poplack & Meechan (1998) later devise other methods of grouping whereby borrowings are largely classified as such using their degree of integration into the recipient language as a heuristic.

Myers-Scotton (1992) takes a different tack, attacking the category of 'nonce borrowing' as a purely methodological concern with little explanatory power. She argues that as products borrowing and codeswitching differ in one key aspect: “[the use of] borrowing is a phenomenon open to monolinguals while codeswitching is not” (1992: 32). Myers-Scotton goes further, noting that borrowing and codeswitching differ in terms of “freedom of occurrence”—essentially appealing to syntactic integration. One argument from her study which merits consideration here is the claim that borrowed forms which are not borrowed due to the introduction of new objects or concepts—recall Hock and Joseph's (1996) discussion of need and prestige as motivations for borrowing—arise initially as instances of code-switching, and that a continuum exists from code-switching to borrowing: when a single- or even multi-word instance of spontaneous codeswitching is taken up and begins to spread throughout a speech community, it gradually becomes more borrowing-like in the process (for further discussion, see Chapter 5).

Myers-Scotton's model parallels the diffusion/transmission distinction, which differentiates between codeswitching, in which the process and product are one and the same, and the use of borrowings, which involves the products of a completed process. Instances of codeswitching are always instances of diffusion—the transfer of material across linguistic systems, and thus bilingualism, broadly defined, is necessarily a prerequisite for codeswitching. Conversely, as Myers-Scotton (1992) correctly notes, borrowings can be (but are not solely) used by monolingual speakers, who have acquired the forms through transmission (after the initial borrowing) and are taking part in this subsequent process. This schema works well as a theoretical distinction, but in the linguistic output of a bilingual speaker, instances of codeswitching and the use of more established borrowings may be impossible to distinguish reliably. To further complicate matters, there exist puzzling instances like German Handy 'cell
phone’, or Beamer ‘projector’ in which the creation might be said to occur with speakers who have extremely limited familiarity with a second language.

**Handy**, a term with ubiquitous use in modern German, is often thought to be an English borrowing—or at least to be like an English borrowing phonemically and orthographically, but it is clearly not a straightforward loan from English. In his discussion of Handy, Onysko (2007: 54) writes that “The true source of German Handy remains obscure although it is possible to detect some relations to an English origin”, noting further that “the reason why Germans say Handy remains opaque.”

Onysko's own model classifies Handy into the category of pseudo-anglicisms, which he takes to exist outside of the larger grouping of language contact phenomena: drawing on a number of sources, Onysko subsumes the various other phenomena discussed to this point under the overall heading of 'language transmission'.

Onysko identifies four primary types of language transmission: the first is borrowing (the classic examples being loanwords like Gang, cool, and Designer) and includes productive uses of borrowing, which Onysko also describes as 'hybrids' (Abendshow 'evening show', Riesen-Events 'huge events', einloggen 'to log on/in'); all of these involve a transfer of both form and meaning from SL to RL—but note that hybrid borrowings could also be created by a monolingual speaker familiar with an established loanword. The second type of transmission is conceptual transmission without SL form, which involves the novel creation of lexical items on the basis of an SL concept (discussed in, e.g., Hock (1991:399) as calquing and identified earlier in Haugen (1950) as the category of loanshifts). An example of this would be a calque like Wolkenkratzer 'skyscraper', lit. 'cloud-scratcher'. The third type of transmission, which Onysko calls interference, involves the semantic broadening of an RL lexical item to match the semantics of a similar SL lexical item—in this way, Ger. realisieren can mean 'to become aware of' on the basis of Eng. 'realize', a broadening from the original German meaning, 'to make something concrete', a definition also shared by the English term. Onysko takes codeswitching, his fourth and final

---

3 Hock (p.c.) suggests a possible origin from 'Handset', a comparatively rarely-used English term referring to the handheld portion of a traditional landline telephone.
5 Onysko's interference should not be understood with the conventional meaning of the term in the fields of psycholinguistics and second language acquisition. Hock (1991: 398) uses the term loan shift to refer to these cases of borrowing. Haugen's (1950) loanshifts include both calques and instances of semantic extension, i.e. Onysko's second and third categories here.
type of transmission, to involve the syntactic embedding of SL material—multiword units—in an RL matrix clause. An example of this type of anglicism comes from a German McDonalds ad: *About this Frühstücksei lachen ja the chickens*, 'Even the chickens are laughing about this breakfast egg'.

As mentioned, according to Onysko, *Handy* falls outside the overarching category of 'language transmission', although it seems a straightforward solution to postulate a transfer of SL form with limited or no transfer of meaning for these pseudo-anglicisms. However, Onysko's final implementation of these categories, where they are mapped onto different types of anglicisms, involves a distinction between *core anglicisms* and *borderline anglicisms*. Perhaps surprisingly, the former category includes pseudo-anglicisms alongside Onysko's categories of borrowing and code-switching (while interference and calquing are classified as borderline). What unifies the pseudo-anglicisms with the former three categories, then, is *formal salience*, which marks these types of transmission as foreign in some orthographic or phonemic way.\(^6\) This distinction is, essentially, a necessary heuristic for analyses which identify anglicisms synchronically without recourse to etymological information—and confers the methodological advantage of (fairly) simple categorization for large-corpus analyses, since without a pre-existing list of instances of these 'invisible' borrowings, they are impossible to distinguish in an automatic fashion from surrounding monolingual text.

To summarize the development of anglicism classification, I present in the following table six examples of language transmission taken from Onysko's (2007) wordlist, along with the terminology used to refer to these phenomena in three sources which propose distinctions along these lines. This table reflects the development of schemata for categorizing language contact phenomena over time, from Haugen's fairly limited (but seminal) treatment of loans to Onysko's schema, which classifies a far wider variety of language contact phenomena.

\(^6\) As with *interference* in the previous note, Onysko's 'formal salience' is used in an informal, theory-neutral way.
### Table 2.1. Comparison of categorization schemes for language contact phenomena.

As comprehensive as it is, one possible concern with Onysko’s (2007) model is that it distinguishes *prima facie* between borrowing and codeswitching by appealing to syntax and the number of words in a stretch of language, both of which, I suggest, can be problematic (recall here the earlier discussion about frequency vs. syntax/length as heuristics for differentiating borrowing and codeswitching). Consider the following three excerpts from the MZEE.com German hip hop forum:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>cool</em></td>
<td>+</td>
<td>+</td>
<td>loanword</td>
<td>borrowing/ loan</td>
<td>borrowing</td>
<td>core</td>
</tr>
<tr>
<td><em>Abendshow</em> 'evening show'</td>
<td>partial</td>
<td>+</td>
<td>loanblend</td>
<td>-</td>
<td>(productive use of) borrowing</td>
<td>core</td>
</tr>
<tr>
<td><em>Wolkenkratzer</em> 'skyscraper'</td>
<td>-</td>
<td>+</td>
<td>loanshift</td>
<td>calque/loan translation</td>
<td>conceptual transmission without SL form</td>
<td>borderline</td>
</tr>
<tr>
<td><em>realisieren</em> 'realize' (new/extended meaning)</td>
<td>-</td>
<td>+</td>
<td>loanshift</td>
<td>loanshift</td>
<td>interference</td>
<td>borderline</td>
</tr>
<tr>
<td><em>About this Frühstücksei lachen ja the chickens</em> 'Even the chickens are laughing about this breakfast egg'</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>code-switching/ code-mixing</td>
<td>code-switching</td>
<td>core</td>
</tr>
<tr>
<td><em>Handy</em> 'cell phone'</td>
<td>+?</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>pseudo-anglicism</td>
<td>core</td>
</tr>
<tr>
<td><em>state of the art</em>⁷</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>borrowed idiom</td>
<td>phrasal anglicism</td>
<td>core</td>
</tr>
</tbody>
</table>

⁷ This example has not yet been introduced; Onysko’s (2007: 278) classification of ‘state of the art’ as a phrasal anglicism is not included in his initial presentation of his classificatory scheme, but is discussed below.
(1) *Album ist sehr geil, features sind auch gut gewählt, track mit [name of artist] und dem anderm da einfach nur killer.*

'Album is very cool, features are also well chosen, track with [name of artist] and the others there simply just killer'

(2) *[name of artist] is baaaaack ...ich höre grade sein Mixtape "[name of album]" ...der Typ killt jeden Beat*

'[name of artist] is baaaaack ...I'm listening to his Mixtape “[name of album]” right now …the guy kills every beat'

(3) *Ach ja da es soviel ich weiss noch keinen [name of artist] Thread gibt frag ich mal hier: Ist [name of album] ein richtiges Album das man im normalen Laden kaufen kann oder ein Mixtape oder was? Thx in Advance*

'Oh yeah because there's still no [name of artist] thread as far as I know I'll ask here: Is [name of album] a proper album that one can buy in normal shops or a mixtape or what? Thx in Advance'

In the first example, we find no variance from the syntax of informal spoken German, nor do we find multiword units in English—under Onysko's model, this could unproblematically be classified as borrowing. In the second and third examples, however, we find entire syntactic units—*[Name of artist] is baaaaack* and *Thx in Advance*—in English, each of which would qualify as codeswitching under Onysko's framework, assuming a straightforward extension of codeswitching to include intersentential switching. However, it could be asserted, depending on one's favored theory of syntax and the lexicon, that these multiword units belong to the class of set phrases or formulaic expressions—and could thus be pulled whole from the lexicon of a monolingual speaker, rather than created in the moment through a compositional process drawing on one or multiple language systems. This, then, is the weakness of this sort of syntax-based classificatory scheme in terms of drawing distinctions between different types of language transmission—determining what exactly is (or isn't) syntax can be a tricky proposition, and Onysko (2007: 277-286) acknowledges this in a discussion of borrowings found in his data, introducing the term *phrasal anglicisms* to cover multiword instances, like *state of the art*, which

---

8 For these excerpts, all material identified as English is presented in bold face in the original and translation. Artist names, song titles, and album titles are omitted, to indicate that they are not relevant to the analysis (as the use of English in these does not imply a conscious choice between languages).
essentially act as borrowings. However, Onysko seems to be one of the few researchers who recognize these multiword borrowings. King (2000: 82), for example, writes, “Intuitively, codeswitches and borrowings would appear distinct: the former is the product of two grammars, the latter of one. In the case of multiword tokens, their classification as codeswitches is straightforward” (2000: 86). Unfortunately, King neglects to give further advice as to how one might intuitively determine the number of grammars at work in a given utterance, and, as noted above, classifying multiword utterances as codeswitches is not entirely fruitful in the case of idioms and formulaic expressions. Poplack & Meechan (1998: 129) put forth a similar sentiment which avoids some of the mentioned pitfalls by relying on the process/product distinction:

[…] codeswitching and borrowing differ as processes, though they may bear some resemblance in surface manifestation […]. Since codeswitching implies alternation between two (or more) language systems, (single-word) codeswitches should show little or no integration into another language. Lexical borrowing, on the other hand, refers to the incorporation of a lexical item from one language into another, with only the recipient system operative.

[emphasis Poplack & Meechan's]

Despite the acknowledgement that codeswitching and borrowing may be difficult to distinguish on a case-by-case basis, Poplack & Meechan (1998) still classify all intra-sentential multiword units as codeswitches without further discussion. An example from Hock (1991) is illuminating here: Poplack & Meechan's scheme would classify the French-influenced plural form courts martial, which is so established as to be the prescriptively preferred form of the noun court martial in English, as codeswitching on two counts: it does not pattern with regular English plural-formation morphology, and it consists of more than one word. However, there is good reason to treat court(s) martial as a reasonably established loanword or borrowing in English, given the word's centuries of use in English (attested in its present form from 1651), rather than as an instance of French-English codeswitching. In addition, Onysko (2007) presents data which reveal an additional shortcoming in taking morphosyntactic integration as a heuristic for the simple distinction of borrowing and code-switching, as Poplack and Meechan do in the above citation: a number of inherent processes of affixation and morphological integration might be blocked by phonotactic or other properties of the borrowed item, i.e., nativization is not a universal process among lexical items, as evidenced by Onysko's (2007: 189) findings: des

---

Crash, des Entertainment, des Laptop, and des Underdog; and des Crashes, des Entertainments, des Laptops, and des Underdogs (the latter group showing overt native genitive marking) appear to occur in free variation in his corpus. Finally, Hock (1991: 413) discusses varying outcomes cross-linguistically in decisions between borrowing and calquing, and he presents the preference of a group of speakers for one or the other as a matter of language ideology—for example, Icelandic (tal/rit) síma, 'wire' for 'telegram, telephone, etc.' Chapter 5 provides plenty of evidence that equally commonly used loanwords likewise vary in outcome with regard to morphosyntactic integration, and as will be seen in Chapter 7, this degree of integration may be conditioned by ideological factors.

Returning to Onysko (2007), it seems that his treatment of formulaic multiword items as phrasal borrowings is the most accurate way of describing these particular phenomena. As with the heuristic of formal salience introduced earlier, however, Onysko espouses a methodological preference for a general scheme reliant on the distinction between morphosyntactically integrated, single-word borrowings and unintegrated, multi-word codeswitches, admitting that “this distinction is not a hard and fast rule but merely holds for the majority of the data.” (2007: 273)

2.4 A basic classification scheme for the study of anglicisms in the present corpora

In light of certain fundamental similarities between Onysko's corpus and my own, specifically the written, publically available nature of the texts, many of his classificatory decisions are relevant in both cases. However, given that the project proposed herein seeks to shed light on the processes of borrowing and due to the methodological concerns deriving from the problem of automatic anglicism identification, I avoid a number of distinctions which have their basis in assumptions about these processes. For reasons outlined in the previous subsection, I similarly avoid distinctions based on degrees of morphosyntactic integration. I take instead a product-based approach which characterizes anglicisms in terms of more accessible formal features. In defining the term anglicism throughout the remainder of this work I will use an adapted version of Onysko's (2007) model, drawing additionally on the insights and terminology of other authors mentioned above and attempting to avoid conflicts with terms used in technical senses elsewhere.
Following Onysko, I unify a number of diverse language-transmission phenomena under the single label of *anglicism*. For present purposes, *anglicisms* as a category include anything used in a German-language context which is, or is generally considered to be, English-derived, i.e., which originates in any variety commonly understood by linguists to be a variety of English.\textsuperscript{10} The category of *anglicisms* is divided into *visible* and *invisible* anglicisms, categories which roughly coincide with Onysko's *core* and *borderline* anglicisms.\textsuperscript{11} *Visible anglicisms*, then, should be understood to include both:

a) *borrowings*, a category which subsumes Onysko's *borrowings* (e.g., *Beef*, 'argument'; *cool*) and instances of Onysko's *codeswitching*, (e.g., *new tracks are up on [name of record label] MySpace page*) including productive uses of borrowing (e.g., *drogendealen*, 'to deal drugs'), phrasal idiomatic multiword expressions (e.g., *in your face*), and nonce and established borrowings as defined by Poplack et al. (1988), regardless of frequency or level of integration, and

b) *pseudo-borrowings*, including items like *Handy*, discussed above.

*Invisible* anglicisms, on the other hand, include:

c) *calques* (e.g., *doppelzeit*, 'doubletime*'), including instances of Onysko's *conceptual transmission without SL-form*; and

d) *semantic extensions* (ex. *frieden*, 'peace' [as a leavetaking], *fett*, 'fat [substance] [as an adj. meaning cool]*), which coincide with Onysko's *interference*.

A number of traditional but contentious distinctions, such as that between borrowing and codeswitching/codemixing, have been obviated by the rather general use of the term *borrowings* here. In the previous section, I have laid out a number of reasons why the existing heuristics for distinguishing code-switching and borrowing are inappropriate; with regard to the data of concern for the present project, of which Examples (1-3) in section 2.3 are typical, as long as the material can be identified as English-derived, the distinction between borrowing and codeswitching is unnecessary. I do not, however, propose that the distinction is universally unwarranted, and it is possible that other researchers may uncover legitimate grounds for proposing such a distinction in their research.

\textsuperscript{10} Examples in the following schema come from the MZEE corpus (see Chapter 3).
\textsuperscript{11} This terminological decision is made due to the various uses of the term *core* with regard to borrowings and the intuitive dissonance of a system which classifies *pseudo*-anglicisms as *core* in some sense.
The heuristics set forth here are advantageous in that they can be implemented computationally in a large corpus, and it should be noted that these are not proposed as deep theoretical distinctions, but rather as practical classifications which are likely to yield reliable results. For reasons of frequency and computational tractability, I will primarily concentrate in the coming chapters on the analysis of visible anglicisms. In the next chapter, I will discuss the collection of the corpora used in this dissertation and begin the application of the definition of anglicism used here to the corpus data.
CHAPTER 3
THE CORPUS DATA

Rap für verblödete Kids
Alle meine Fans fliegen von der Schule
von der Sonderschule
und gehen dann jobben in 'ner Pommesbude

Rap for stultified kids
All of my fans flunk out of school
out of special school
and go and get jobs at a fast food restaurant

–Favorite, 'Alle Scheisse', 2011

3.1 Introduction: Forums as corpora

The data analyzed in this dissertation include two corpora I collected for the purpose from online discussion forums. In this chapter, I will introduce forums as a type of computer-mediated communication or computer-mediated discourse (Herring 2007) and discuss the transformations whereby the Internet forum pages are translated to their corpus format, which makes use of XML (eXtensible Markup Language). Next, I will describe the two specific forums from which data in this dissertation are drawn, presenting a preliminary by-hand analysis of anglicism frequency and function in the German corpus.

Internet forums are a form of computer-mediated communication involving a number of registered users who can interact with the forum in two primary ways: by posting a new thread, which, like an e-mail, usually involves a shorter subject followed by a longer message; or by posting a message replying to an existing thread. The individual messages on a forum are referred to as 'posts', and may be conceived of as similar to an e-mail message sent to a group or list. A primary difference is that these posts are collected and displayed on the Web, and forum users must return to a specific website to view the posts. Another difference is that forum threads are usually public, and in general, can be viewed by all registered users, or, commonly, even unregistered users. Users are identified by a username, which can effectively provide anonymity. Threads are generally listed within a forum by subject, with the thread with most recent activity appearing highest on the page. When the thread subject is clicked, the viewer is taken to the

12 The characterization of forums presented here is not universally applicable; for practical purposes, I characterize the type of forum which is most commonly found on the English-speaking and German-speaking portions of the Web, and more specifically the type of forum from which the present data are taken.
thread in its entirety (or, if lengthy, divided into several pages), with replies commonly listed from top to bottom in chronological order. Users who have a particular interest in a thread may have the option to subscribe to the thread, receiving e-mailed alerts when someone posts to that thread, or may choose to be notified via e-mail whenever someone posts to a thread in which they have posted. In addition, users who post to a thread may have the ability to automatically quote all or part of a previous post with an automatic indication that the reproduced material is a citation.

In addition to registered users, forums commonly have several other participant classifications: guests, or unregistered users, who generally have permission only to view the forum (but cannot post or reply to threads), moderators, who have the ability to delete or edit threads and restrict or ban users from the forum, and administrators (usually those who own or rent the server space), who have the abilities of moderators as well as the ability to promote users to moderators or demote moderators to users and set forum rules and policies. Larger forums are often divided into subforums by general topic category; administrators have control over the site in its entirety, while moderators can often perform limited actions in specific subforums. One characteristic which is especially relevant to linguistic analysis is the persistence of forums; unless posts are deleted by an administrator, moderator, or automated script, forum threads are generally persistent, in that they are often indefinitely accessible. To reduce the number of threads displayed on the main page, however, forum threads over a certain age are often only accessible through an archive link at the bottom of the page.

3.2 Corpus collection

The forums from which I have collected data for the present project are similar in that they operate (on the server side) on the popular forum software vBulletin, which keeps an automatic archive including all threads posted to the forum since its inception, usually in a format closer to plaintext (omitting, e.g., users' chosen avatars or icons, rank, etc.) than the default forum view, and displaying each thread in its entirety on a single page. Figure 3.1 shows several posts in normal view and Figure 3.2 shows the same thread in archive view to illustrate the difference.
The archive view, in terms of its underlying HTML (Hypertext Markup Language) format, contains much less extraneous information while reproducing the written text exactly; the primary drawback is the lack of differentiation between original and quoted text (the latter can be seen in the speech bubble in Figure 3.1):

Figure 3.1. Posts from a German forum, normal view.

Figure 3.2. Posts from a German forum, archive view.
Using a *spider*, or automatic script, it is possible to automatically collect the HTML files from all or part of a forum, although for larger collections, measures to reduce server impact should be put into place.\(^{13}\) In order to build the forum corpora used here, I use a single script to gather a list of URLs for individual threads, collect HTML files from the archive and re-tag them in XML format, concatenating them into a single file. Specifically, the following information is saved as part of this process: Each thread is enclosed in `<thread></thread>` tags including the original thread number. Posts within threads are enclosed in `<post></post>` tags, and are numbered in chronological order, each with a unique number. The poster's username (replaced with a unique number so that posts from the same user are still identifiable), and the date/time of the forum post are similarly XML-tagged. The main text of the post is enclosed in `<text></text>` tags.

3.3 Corpus normalization and quoted/extraneous material

In order to facilitate the further processing of the forum data, several initial transformations were made to the corpora. All of these changes were guided by the principle of changing the forum users' natural and original linguistic output as minimally as possible while making the XML corpus both human-readable and accessible to script-based text processing.

The first concern was a collection of issues revolving around the display of text: punctuation marks and special characters encoded in UTF-8 and HTML (e.g., `&quot;` & `&amp;` & `&gt` & `&lt`) were replaced with their Latin-1 (ISO-8859-1) equivalents, and stray HTML tags and bulletin board codes (largely artifacts of previous versions of the forum hosting software, and occurring in only a small portion of the corpus) were removed.

The second concern was the separation of quoted or cited material from original text. This is desirable because, especially in terms of measuring word frequency, text which has been quoted or cited is not essentially original linguistic production. For example, a unique anglicism contributed by a single user may be reproduced through quoting and citation by other users without alteration, artificially inflating its apparent frequency in a simple corpus search. For this

\(^{13}\) The spider used for the current project, which additionally provided the initial conversion to XML format (and which used a timed delay between collections to reduce server impact) was originally developed by Liam Moran, who I would like to thank here for this contribution. I made subsequent modifications to this script in order to collect and convert additional corpora from other sources.
reason, I automatically tagged words or multi-word strings enclosed in single or double quotes in <quote></quote> tags, which occupied the same level in XML structure as the <text> tags, i.e., the level below <post>. As noted above, material automatically cited or quoted using the forum's built-in affordance for this purpose was not differentiated from original text in the archive view. For this reason, I also implemented an algorithm to tag lines identical to any of the previous 200 lines within the same thread as quoted material, as the vast majority of forum citations occur in reasonably close proximity to the original posts. Other questionably linguistic material was then separated from the main text with <junk></junk> tags, including e-mail addresses, URLs, and emoticons (specifically those represented in the archived text using vBulletin's conventions, e.g., :mad: for an angry emoticon; these may be typed in by users, but are more commonly added to the text through menu buttons in the post-composing interface). An example post (containing original text, quoted text, and an emoticon) in the XML corpus format is reproduced in (1a) and translated to English in (1b) below. In this case, only the material between the <text> and </text> tags is considered an original linguistic contribution; the material between the <quote></quote> and <junk></junk> tags would be ignored for any non-qualitative analysis:

(1a)  
<post number="000310">  
  <username>000130</username>  
  <date>02-11-2005, 21:36</date>  
  <quote>  
  bist du der robin hood?  
  </quote>  
  <text>  
  ja genau, du bist ja richtig der schlaue junge  
  </text>  
  <junk>  
  :rolleyes:  
  </junk>  
</post>  

(1b)  
<post number="000310">  
  <username>000130</username>  
  <date>02-11-2005, 21:36</date>  
  <quote>  
  are you robin hood?  
  </quote>  
  <text>  
  yeah exactly, aren't you a clever boy  
  </text>  
  <junk>  
  :rolleyes:  
  </junk>  
</post>
The decision to separate quotes and extraneous material from the main text, rather than to delete them from the corpus, was made in order to preserve the context of original linguistic production when, e.g., performing by-hand qualitative analysis. The preliminary text-processing steps above were performed for both forum corpora used in this dissertation; these corpora and their sources are the focus of the next two sections.

3.4 The German hip hop forum: *Hip Hop Diskussion* at MZEE.com

The primary corpus is drawn from the forums at the website MZEE.com, which presents itself as the größtes deutschsprachiges Hip Hop medium überhaupt 'largest German-language Hip Hop medium anywhere.' Independent web monitoring company Alexa.com shows this claim, in terms of online audience, to have been valid through the first quarter of 2009.14 Today, MZEE is a prominent online hip hop clothing and record retailer, hosting not only forums but also blogs and an event calendar. The MZEE.com forums, with roughly five times as many posts but also blogs and an event calendar. The MZEE.com forums, with roughly five times as many posts archived as its closest competitor, include debate about German and American hip hop artists and culture alongside original lyrics and off-topic general discussion threads.15 Participants on these forums are German-speaking hip hop fans volunteering their views and providing discussion on topics germane to hip hop culture; as of May 2012, Alexa reports that 84% of visits to the site originated in Germany, with roughly 3% originating from each of Switzerland and Austria, and 2% of visits originating from the US; these four countries, then, jointly contribute roughly 92% of site traffic, with other countries contributing less than 2% each.

In two collection instances in 2009 and 2011, I collected a total of 12,540,944 word tokens in 381,880 posts (after the above-mentioned processing) from the largest forum and the first forum listed among the MZEE forums. This forum is entitled *Hip Hop Diskussion*, which has the tagline Hier wird alles, was mit Hip Hop zu tun hat diskutiert, 'This is where everything hip hop-

15 As of 13 May 2012, mzee.com featured 5.25 million posts and 108,000 registered users; hiphop.de featured 1.04 million posts and an unknown number of users; and rap.de featured 0.54 million posts and 54,000 registered users. Note that these numbers indicate the number of posts available as of this date; the rap.de forums, for example, had over a million posts when checked in 2009; it is likely that some have since been removed to save server space.
related is discussed.' The MZEE corpus, a term I will use throughout this dissertation to refer to this collection, essentially mirrors the archives of the entire *Hip hop Diskussion* forum at MZEE.com from March 2000 to March 2011, a span of time which makes diachronic analyses possible. It is worth noting at this juncture that it is nearly impossible to be sure that all users of this forum are native German speakers, or that some are furthermore not native English speakers. While such variability certainly exists in the data, the large size of this corpus and the likelihood that the vast majority of these forum users are native speakers of some variety of German (see, e.g., the Alexa traffic by country given above) should serve to mitigate some of these concerns. At the very least, it can be asserted that this dataset represents speakers of German, and furthermore speakers of German who are interested in hip hop. Figure 3.3 shows the MZEE.com forum frontpage:

![Figure 3.3. Frontpage of the MZEE.com hip hop forums, 12 May 2012.](image)

In order to provide certain points of comparison with the MZEE corpus, a maximally similar corpus was collected from an English-language online hip hop forum based in the United States. This corpus is described in the next subsection.
3.5 The American hip hop forum: Project Covo

Project Covo is the name used for the forums hosted at SOHH.com, an English-language hip hop news website founded in 1996 (SOHH being an initialism for Support Online Hip Hop). In 2011, I collected a total of 19,385,022 word tokens (after processing) in 610,450 posts from the 'The Spot' subforum on Project Covo, which is billed as 'The place to be for general discussions on Hip-Hop,' and constitutes the first-listed and largest forum on the site in number of posts (with almost 4 million available as of 2012). This corpus includes posts from June 2003 through November 2011. Alexa reports that 64% of traffic on the site originates from the United States, with 10% coming from Bermuda, 6% from the UK, 3.5% from India, and less than 3% each from all other countries. Figure 3.4 shows the frontpage of the Project Covo forums:

![Project Covo forum frontpage, 12 May 2012.](image)

The Covo corpus is, in general, comparable to the MZEE corpus in discussion topic, purpose, format, and timeframe; the primary differences between the two are overall size (19.4 million vs. 12.5 million word tokens, but this difference is neutralized by relativizing word frequency) and the primary language of discussion. This second difference enables the investigation of anglicism frequencies in the MZEE corpus with respect to the incidence of the original English words in the Covo corpus, which can show the time difference between the popularization of a term in English and its later adoption as an anglicism in German; such an analysis is presented in Chapter 6. In the next section, I provide a by-hand analysis of a subset of the MZEE corpus in
order to determine the type of anglicisms found and the general frequency with which anglicism tokens and types appear in the text.

3.6 First steps: Anglicism frequency in the MZEE corpus

The next chapter focuses on the development of a computational classifier which automatically identifies anglicisms in the MZEE corpus. This classifier provides a means of identifying a large number of candidate anglicisms, but, in the end, provides such a long list of candidates that it is not feasible to hand-correct the entire output. Using the classifier, however, I am able to identify anglicism wordforms which occur more than one hundred times in the MZEE corpus. That result is a crucial step which opens the door to more advanced computational analyses, but it does not provide a key piece of information, namely the overall incidence of anglicisms in the text. In this section, I present a preliminary by-hand analysis of anglicism frequency in a subcorpus—roughly 24,000 word tokens (and 1000 posts) in length, including 100 posts drawn from each of 10 MZEE threads. This by-hand analysis provides two major benefits: it establishes a baseline frequency for anglicisms in the text—the percentage of the text that is English-derived, and is useful to establish a ‘gold standard’ for assessing the performance of the parser described in the next chapter—and it affords an in-depth engagement with the corpus. The by-hand identification task allowed me to begin to catalog the kinds of anglicisms present in the corpus and how those anglicisms were nativized, for example, by acquiring German inflectional or derivational morphology, or by compounding with German words. These insights in turn led to invaluable advances in the development of the computational identification system.

The 10-thread subcorpus analyzed here includes data from the years 2007 and 2008 from threads discussing a number of German and American hip hop artists. These threads were selected at random from a list of notable artist discussion threads compiled by the users of the MZEE.com forum. In terms of identifying anglicisms in the corpus, I have the qualifications of being a native speaker of English and a fluent German speaker and reader, with eight years of German instruction, two years of living in Germany, and several years of continuing practice with the language. In addition, I am familiar with English and German hip hop culture, having listened extensively to both English- and German-language hip hop for almost a decade and a half, as
well as co-hosting a radio show featuring German, American, and Japanese hip hop music as an undergraduate student.

Subjectivity in this kind of human-based language identification task is on some level inescapable, but is mitigated in this case by the use of a reliable online bilingual dictionary (LEO GmbH 2008). If the candidate anglicism was found in the dictionary with the relevant meaning/translation into English, it was considered an established anglicism—i.e., one which was likely in general usage before hip hop culture became influential in Germany, and thus of reduced relevance to the current study. Examples of such anglicisms include items like Baby, Party, and cool, which have long been established as part of standard German usage. In the identification process, certain parts of the text were exempted from anglicism identification, including quoted song lyrics, names of hip hop artists, album and song titles, names of record labels, forum posters' usernames, names of TV shows, magazines, and movies, and direct quotes from previous posts and outside news sources—in essence, the kinds of material which were considered separately when performing preliminary text-processing on the corpora. These were exempted from identification because no real choice is made by the speaker to favor one item over the other: these, then, can be considered a type of forced usage.

Once anglicisms in the subcorpus were identified, they were grouped into stem types: all occurrences of declined verbal, nominal, and adjectival forms of a single lexeme were taken to constitute a type. At this point, Carstensen & Busse's (1993) Anglizismen-Wörterbuch 'Dictionary of Anglicisms' was consulted for matches. This reference was chosen primarily because it is the most complete dictionary of anglicisms in German from 1945 on, but it also has the particular benefit (with regard to this study) of its year of publication. Carstensen & Busse's corpora extend up to about 1989, when hip hop in Germany had yet to appear significantly in the cultural radar of the mainstream. Therefore Carstensen & Busse's dictionary is unlikely to include most newer hip hop-related borrowings, assisting in the proper identification of these

---

16 'The term 'type' here is used in this sense (which could be considered more similar to 'lemma' or, even more accurately, 'stem-containing wordform') in the service of maintaining terminological consistency in order to facilitate comparisons with Onysko's (2007) similar analysis. In the next few chapters, I use the term 'type' in, e.g., 'accuracy over types' in a different way, namely to indicate that all instances of a single wordform (rappers being a distinct wordform from rapper) are considered together. When the distinction is relevant, I will use 'stem type' to indicate the former and 'wordform type' to indicate the latter.
borrowings through removal of older borrowings from consideration. Those anglicisms which were found to match in form and general meaning (including straightforward extensions, like *crew*, attested from 1910 with the nautical meaning) in Carstensen & Busse's dictionary (423/611 types, or 69%) were exempted from consideration, as these anglicisms are not candidates for hip-hop related borrowing/innovation. The remaining 188 types were categorized into one of three kinds of novel anglicisms, the first being hip hop-related borrowings, which are group-exclusive or nearly group-exclusive to the community of hip hop fans, and which carry particular social meaning relevant to that community (e.g., *yo, Beef, word*). The second category of anglicisms not found in Carstensen & Busse (1993) is comprised of borrowings related to the Internet and other new technologies which appeared post-1993—some of these refer to technological innovations (e.g., *Download/downloaden, homepage*), while others reference engagement with a shared Internet culture (e.g., *lol*). The third and final category of novel anglicisms found in the by-hand analysis involves borrowings related to some aspect of the music industry, but which are not exclusively (but may be preferentially) used in discussions of hip hop (*Release/releasen, lyrics, Producer*). Notably, *Rap, Hip Hop, rappen, and Hiphopper*, along with their orthographic variations, were removed from this analysis, as all of these terms appear in Carstensen & Busse (1993). Although these terms are certainly used within the hip hop community, they are also used in broader societal discourse, by Germans who do not necessarily affiliate with hip hop culture—these terms, then, are not group-exclusive. This is not to say that all of the anglicisms captured in the by-hand analysis were intended to be group-exclusive; the classification of identified terms with recourse to the dictionary was intended only to ensure that the anglicisms categorized in this analysis were not already well-established loanwords.

The examination of these novel anglicisms found in the by-hand analysis provides a first look at the sources of English influence we expect to find in this domain. The data are not only distinctive because they constitute language use in hip hop culture, but also because of their status as computer-mediated discourse, which is important to keep in mind in considering the later analyses. In addition, it seems that novel anglicisms exist, to some extent, on a continuum of group-exclusivity—anglicisms from the first category, which tend to associate their users strongly with hip hop culture, might be expected to be more group-exclusive than those that are—in essence—terms of art from the third category, such as *Producer*, that could be expected
to be relevant to multiple musical cultures. However, the use of *Producer* rather than the longerestablished *Produzent* (an earlier borrowing from French) might be considered a mark of group membership and even ideological orientation. These issues will be explored in greater depth in the remainder of this dissertation.

I turn now to the presentation of descriptive statistics for the 10-thread MZEE subcorpus, comparing the results to a similar analysis in Onysko (2007). In a quantitative type/token analysis, Onysko (2007: 114) found the proportion of anglicism stem types to overall word stem types in the 2000 volume of German newsmagazine *Der Spiegel* to be 5.8%. In the 10-thread corpus analyzed here, the proportion of anglicism types to word types overall was found to be slightly higher (6.89%) as displayed in Table 3.1.

The most dramatic result of this analysis is that the present findings differ greatly from Onysko's findings for the 2000 volume of *Der Spiegel* in terms of the proportion of anglicism tokens to wordform tokens overall, taking into account the number of instances of a single word. Onysko (2007: 114) found a proportion of 1.11% anglicism tokens to word tokens (straightforwardly, 1.11% of the words in the corpus were anglicisms), but this figure is several times higher, at 7.35%, in the MZEE subcorpus analyzed here.

<table>
<thead>
<tr>
<th></th>
<th>Stem types</th>
<th>Wordform tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of words, 10-thread corpus</td>
<td>4454</td>
<td>23907</td>
</tr>
<tr>
<td>Number of anglicisms</td>
<td>307</td>
<td>1755</td>
</tr>
<tr>
<td>Percentage anglicisms</td>
<td>6.89%</td>
<td>7.35%</td>
</tr>
</tbody>
</table>

*Table 3.1. Incidence of anglicisms (new and old) overall in the 10-thread corpus.*

17 In this case, the distinction between novel and established anglicisms is discarded; the overall incidence of all anglicisms in the subcorpus is used in this comparison, as no distinction was made in Onysko's results.
What these two proportions tell us with respect to Onysko’s findings is that, while proportion of borrowed words from English used in the MZEE.com forum is only slightly higher (6.89%) than in *Der Spiegel 2000* (5.8%), these anglicisms are used much more frequently in the case of the MZEE.com forum (7.35%) than in *Der Spiegel 2000* (1.11%). This finding underlines one of the primary reasons that the domain of hip hop was chosen for analyzing anglicisms—namely, that English-derived forms are especially frequent in this domain. In addition, a great number of the anglicisms used in this domain are recent, unestablished forms that stand to illuminate the particular processes involved in language contact and language change.

3.7 Summary

In this chapter, I described the properties of Internet forums as a type of computer-mediated communication and introduced the MZEE and Covo corpora collected for this research project, discussing preliminary operations undertaken to normalize the text while retaining as faithfully as possible the original natural language contributions of forum users. I also presented a preliminary, by-hand analysis of anglicisms in a roughly 24,000 word subcorpus excerpted from the MZEE corpus in order to establish the observed relative frequency of borrowed lexical items, which, as Gries (2010) notes, is a necessary metric for comparing frequencies across texts. The by-hand analysis also provided exploratory insights into the varied functions of anglicism forms and the morphological adaptation of these forms. These latter findings will be discussed as they become relevant in the next chapter, where I describe the development of a computational system for automatically identifying anglicism wordforms in the MZEE corpus.
4.1 Introduction

In terms of the overarching research questions which this dissertation engages (see section 1.3), the large-scale corpus work presented in this chapter primarily addresses the automatic identification of anglicisms in text. However, several other issues are also brought to the fore in this chapter—in constructing a computational system to identify anglicisms, it is necessary to apply knowledge about the ways in which anglicisms are nativized, particularly in terms of acquired morphology and compounding with German words. In addition, the results of the classification task reveal which anglicisms are most frequent overall in the MZEE corpus.

4.1.1 Motivation

The primary concern in this chapter is the construction of a reliable classifier—a computational tool or system which can, automatically and without intervention, predict the category of new units of data (categorizing, for example, documents, sentences, or words) based on a previously-categorized set of training data. In the present study, the main goal is the development of a binary classifier which can categorize orthographic words as either German (labeled G) or as possible English-derived forms (labeled E). Although the amount of effort involved in the development of such a classifier is significant, it is worthwhile as it provides a number of crucial advantages in large-scale corpus analysis. In the present study, anglicisms are distributed relatively sparsely.

---

Yo, der Beat erlaubt's, ich attackier mit smoothem Rap,
Junge träum schön weiter vom Plattendeal bei Groove Attack.
Ich appellier für 'nen guten Zweck an die wahre Szene,
und plakatier: "Steckt Wanna-bes in Quarantäne!"

Yo, the beat allows it, I'm attacking with smooth rap
Kid, just keep dreaming of a record deal with Groove Attack
I make an appeal for a good cause to the true scene
and put up posters: “Put wanna-bes under quarantine!”

– Nico Suave on 'Session' by Samy Deluxe, also featuring Dendemann & Illo 77, 2001

---

18 This chapter includes some material to appear in: Garley, Matt & Julia Hockenmaier. Forthcoming, 2012. Beefmoves: Dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. Proceedings of the 50th Association for Computational Linguistics Short Papers. I would like to particularly acknowledge Dr. Hockenmaier’s contribution to the development of the compound-cutting algorithm described in this chapter.
throughout a large body of text, and the correct identification of a large number of these through by-hand analysis quickly becomes prohibitive. Without an automatic means of identifying anglicisms in the text, analysis would be limited to a relatively small set of words chosen by the researcher, which could prove detrimental to the accurate characterization of anglicisms.

The process of developing and refining a reliable classifier for the present dataset also serves to provide additional information about the linguistic properties of anglicisms in the text. As an example, a fairly basic classifier, given the past participle *geflowt*, 'flowed', may label the word as German, but the removal of the German affixes *ge-* and *-t* from the word upon input to the classifier (as covered in Section 4.3.2) improves the classifier's performance considerably; this, then, provides the researcher with confirmation that these affixes are regularly added to borrowed English words. If the removal of such an affix fails to change the classifier's performance or proves detrimental, the researcher learns that this affix may not be used with English borrowings. In devising numerous ways to improve the performance of a classifier, a large amount of information about the morphological behavior of borrowings can be gleaned.

Finally, the development of a reliable multilingual classifier is a non-trivial problem in the field of natural language processing. While the classifier described here is specialized in its intended application, the methods used to improve the classifier's output are in several cases novel. Future research on the development of specialized or customized classifiers for similar multilingual data stands to benefit from a detailed description of the methods used to improve the present classifier's output, and while the system as a whole is specialized, many of the individual methods used in its development are of more general interest.

4.1.2 Related work

A related task is described by Alex (2008), who constructs an 'English inclusion classifier' for multilingual texts. While the system presented here for the automatic detection of anglicisms takes a number of cues from Alex's English inclusion classifier, the overall goals of the present research are significantly different, and the details of each system's development and performance vary considerably. The anglicism classifier presented here is intended to reliably identify novel (i.e., not previously known, not widespread or well-established) and, particularly,
hip hop-related anglicisms in order to examine their features, nativization, and use. By contrast, Alex’s English inclusion classifier seeks to identify English words for the purposes of natural language processing in a more general application—as a result, choices were made in preparing hand-coded data for the English inclusion classifier which would not be appropriate in the present research. To give several examples, Alex’s (2008: 47-53) system:

- sought to capture 'foreign words' in general, including internationalisms (and homographic words)
- sought to include some 'names of English origin', including some placenames, but not personal names (*Hutchison, Forrester*)
- ignored hybrid/bilingual compounds during evaluation (*Shuttleflug, 'shuttle-flight'*)
- ignored English words with German morphology during evaluation (*Receivern, 'receivers'*)

Because the target of the present research is to capture English language material, other foreign words and internationalisms are beyond its scope. Proper names of English origin, like placenames and the names of hip hop artists, are also not a major concern here because these do not constitute a choice on the part of the speaker/writer, and as such they do not yield the sociolinguistic insight that other words do. The issues of loanword compounding and the use of German morphology with anglicisms is a major departure from Alex's goals; which words are nativized and how that process works is a major research question engaged in this chapter.

4.2 Methodology of classification

The primary task presented in this section encompassed the automatic detection of novel anglicisms in a German-language hip hop discussion forum corpus, i.e., the MZEE corpus presented in Chapter 3. In order to prepare this corpus for the present work, a new version was created as the result of several operations (in addition to the corpus cleaning operations presented in the previous chapter). This 'list' version of the corpus excluded previously-identified quoted and extraneous material. Words were defined by surrounding whitespace, including spaces, tabs, and line breaks. Non-alphabetic characters were treated as follows: `<β>`, German 'double s', was
replaced with `<ss>`, and the following symbols were treated as word breaks: exclamation point, at symbol, percent, ampersand, parentheses, brackets, curly braces, colon, semicolon, comma, period, tab, newline, question mark, forward and backslash, hyphen, underscore, plus, equals sign, and the tilde. All other non-alphabetic characters and numbers were removed from words.

At its core, the automatic detection of anglicisms in a large corpus is a binary word classification task, i.e., a task in which individual words are labeled as belonging to one of two categories according to a certain algorithm. For this task, I chose to use MALLET (MAchine Language for LEarning Toolkit) (McCallum 2002), which among its other features contains a classification module. For present purposes, MALLET works as follows: the researcher prepares a set of training data with each word labeled, e.g., a German word could be labeled G. These training data (and any data to be classified) then undergo a process of feature extraction; ‘features’ in this sense having a very flexible definition involving any of a set of attributes selected by the researcher that are associated with the word. As an example, a very basic feature extraction could take the orthographic word `<dog>` and extract each character as a feature, yielding the feature-frequency pairs ‘d:1 o:1 g:1’, as `<dog>` has three letters, and each occurs once. The word `<deeded>` would thus be represented by the features ‘d:3 e:3’. The set of feature-frequency pairs for each word is known as a feature vector. Based on the labels and the feature vectors in the training data, MALLET produces a classifier using one of several algorithms. In the present study, a maximum entropy algorithm was used. Creating and using such a classifier involves the following steps. First, appropriate training data are selected and labeled. A development set, in this case, 10,000 tokens of running text from the MZEE corpus, is coded by hand (functioning as a ‘gold standard’). Desired features and corresponding frequencies are then extracted from the training data and development/test sets. The classifier is trained using a series of commands to the MALLET package, and the classifier is then used to classify the development set. Through the analysis of errors on the development set, features are refined, problems with the training data are identified, a list of stop words (words to ignore) is identified, and a series of incrementally improved classifiers is created. Through this process, the classifiers can be

---

19 Angle brackets `<>` are used when relevant to indicate orthographic strings (as opposed to, e.g., phonological or phonetic representation or discussion of language material in the abstract.
20 `ä`, `ö`, `ü` and their uppercase variants were included here as ‘alphabetic’.
optimized (using the development set) on several metrics which can be considered extensions of accuracy.

4.2.1 Training Data

Once the classifier is produced, it can label unlabeled data based on the probabilities associated with each word's features. The performance of the classifier is dependent on the type and quality of its training data and the features selected for training and evaluation, along with any post-processing. The production of a classifier is an instance of supervised machine learning. In this case, training data constitutes the extent of the researcher's supervision; this set of accurately-labeled forms, and the feature-frequency pairs for each form, provide the basis upon which the classifier assigns labels to new data. As hand-annotation of a large set of training data is an expensive task, I follow Alex (2008) in opting for the pre-existing CELEX corpora (Baayen et al. 1995), which are distributed in several versions with different attributes. For the present study, the 'German Orthographic Wordforms' and 'English Orthographic Wordforms' versions of the CELEX corpora were selected—these consist of one entry per wordform, including inflected and derived forms, e.g., not only plan but also planned and planning, and as such, are more useful than wordlists containing only lemmas. In addition, wordforms from the Project Covo forum, introduced in Chapter 3, were prepared in list form and used as training data, with the entire corpus labeled E due to the very low probability that an American hip hop forum would contain German-language content. This departure from using only pre-existing corpora was deemed necessary due to the probability of encountering hip hop-specific vocabulary in the data to be classified.

Because of the difference in size between the German and English CELEX wordlists and the small size of the English Project Covo wordlist, the initial training data was 411,335 words in size, of which 87,352 (21.2%) were English, and 323,983 (78.8%) were German. Because of some overlap between the English CELEX wordlist and the Project Covo wordlist, the training data ultimately contained 398,426 unique word-label pairs, of which 74,443 (18.7%) were labeled E.

21 A wordform is defined here as an orthographic string (sequence of characters) along with the linguistic material the string represents.
Table 4.1. Size and composition of training data sources.

Because of the overlap between the source wordlists for the training data, however, it was determined that any word occurring in the training data with conflicting labels (i.e., any word that was labeled both E and G at different points in the data) should be removed. This operation resulted in the removal of 4,340 words (that is, 8,680 word-label pairs) from the training data, for a final size of 389,746 unique words, of which 70,102 (18.0%) were labeled E.

4.2.2 Development

10,000 words of running text from the MZEE corpus were selected for the creation of a development set. After the first 5,000 were hand-coded, early tests indicated that the inclusion of upper/lower-case information had negligible effects on the classifier's performance (detailed results of these early tests are presented in section 4.5.1). This is likely due to the fact that the target data, being collected from an online discussion forum, were produced adhering to few norms related to consistent use of upper/lower case; in addition, because the classifier as used here treats upper-case and lower-case letters as fundamentally different, e.g., <C> and <c> have no greater relationship than <a> and <c>, there is a risk that the associations between letters or strings of letters and their probable labels might be weakened. As an example, if the training data hypothetically contained only ten words beginning with <b>, and all were German words, it is possible that some—say, half, would be nouns, and thus capitalized according to German written norms (this example, of course, leaves aside the possibility of sentence-initial capitalization, or the possibility that forum users largely ignore orthographic norms). When the classifier is required to label a novel word beginning with <b> or <B>, it has only seen five instances of each, and this hypothetical rule (that only German words begin with <b> or <B>) is only half as

---

22 For a more precise characterization of the texts included in the CELEX corpora, see the CELEX readme document, [http://www.ldc.upenn.edu/Catalog/readme_files/celex.readme.html]. (Accessed 30 July 2011)
prominent as it should be in the classifier's computation of probability. A possible solution to this problem would be to include all words upper- and lower-cased in the training data, but, compared to simply lower-casing all words, this should lead to equivalent results on novel data—the training data would just be twice as long. For reasons of simplicity and efficiency, all training and development data were lower-cased.

The 10,000 word development set was annotated with binary labels (E or G) by the researcher according to the following guidelines: wholly German words were labeled G, as were all placenames, proper nouns, artist names (whether German or English), international borrowings (like Amateur), and well-established borrowings (like Baby or Job). Many of these words were also exempted from the classifier's evaluation (see section 4.2.4). Borrowings suspected to be well-established were looked up in an online bilingual dictionary (LEO GmbH 2012) and labeled G if they appeared there as German forms.

English roots with German morphology, compounds containing English material, and 'surface' English lexical content (thus excepting calques or loan translations) are labeled E if they are not well-established borrowings. As an example of the above distinction, Freestyle, Peace, MC, Cash, down, high, and Stage are not found in the dict.leo.org database as German terms and are considered English terms, but more established borrowings like Image, Partner, Rap, Rapper, Interview, and Song are labeled G due to their occurrence in the German dictionary at leo.org. It is worth noting at this point that there are no innate properties of words like the latter that make them significantly different (or particularly automatically identifiable as different) from words like those in the former set—all are words of English origin. However, while this decision necessarily reduces the performance of the classifier—or, more specifically, the ultimate evaluation of the classifier's performance, the entire point of this exercise is to find the specific kind of anglicisms that are relevant to the overall research project.

4.2.3 Features and measures of success

The success of a classifier depends not only on the selection of training data, but on the features selected for classification. These features determine what the classifier 'knows' about any given word. The probabilistic association of each feature and combination of features or feature vector
with a particular label is central to the classification problem. The maximum entropy model used by MALLET creates a deterministic classifier based on the labels and associated feature vectors in the training data. Given an unknown feature vector, this classifier produces a label for a new word. For a more in-depth discussion of the application of maximum entropy models for text classification, see Nigam et al. (1999).

For the present research, a number of different classifiers were created by extracting different feature vectors from the training data. In evaluating these classifiers against the gold-standard coding of the development set, a baseline or naive system is generally used to provide a point of reference. In this case, there are two possible baselines to consider. The most naive sort of system would be one which classifies all tokens with one label. In the present study, classifying all words in the development set as G would achieve an accuracy of 84.5%, given that 15.5% of the words in the development set are labeled E (after stop words and words of 2 characters or less are removed), providing a fairly high baseline. However, this misses all anglicisms, and as such, would be entirely useless given the goals of the present research. In addition to accuracy, the proportion of words in the development set which are correctly classified, there are three other metrics to consider when evaluating classifiers for the present research: precision, recall, and F1.

For present purposes, recall, precision, and F1 are considered only for the label E. Remember that accuracy is the proportion of words (G and E) labeled correctly to the total number of words under consideration. Recall is, in this case, the proportion of words correctly labeled as E by the classifier to the total number of words labeled as E in the gold standard set. Precision, on the other hand, is the proportion of words correctly labeled as E by the classifier to the total number of words (correctly or incorrectly) labeled as E by the classifier. Note that the naive baseline described above would have an undefined (divide by zero) precision on E, and a recall of 0 on E. Both precision and recall are expressed as percentages, with 100 percent being the ideal outcome in each case.

When considered in terms of true/false positives and negatives, we can use $tp$ to denote the number of true positives (words classified as E which are truly E), and $fp$ to denote the number of false positives (words classified as E which are in fact G); likewise, $tn$ denotes the number of
true negatives (words classified as G which are truly G), and \( fn \) the number of false negatives (words classified as G which are truly E) In this case, recall is determined by the formula \( tp / (tp + fn) \) and precision by \( tp / (tp + fp) \) (van Rijsbergen 1979).

F1, the most commonly used F-measure, is the harmonic mean of precision and recall, yielded by the equation \( F1 = 2 \times ((\text{precision} \times \text{recall}) / (\text{precision} + \text{recall})) \). In F1, precision and recall are weighted equally, and as such, F1 can be considered a composite score encompassing the two. Because in the present research I want to capture as many anglicisms as possible with the understanding that the final list will have to be weeded of false positives by hand, recall is more important than precision in this case. However, I will report F1 rather than an alternative F-measure because of its common use in the field.

4.2.4 Evaluation

All classifiers trained using MALLET in the present research make use of \( n \)-gram features. Beginning with 1-grams, as in the earlier examples of 'deeded' as 'd:3 e:3', a feature vector can be built using 1- and 2-grams, so that the feature vector of 'deeded' becomes 'd:3 e:3 de:2 ee:1 ed:2'. Classifiers trained on \( n \)-gram features tend to perform poorly when labeling short words; a 1-through 5-gram classifier (that is, a classifier considering strings of one, two, three, four, and five characters as features) can only perform as well as a 1-through 2-gram classifier on a two-letter word; such words are feature-poor. In addition, most 2-letter words are unimportant for the current research as there are few two-letter anglicisms in the text; exceptions (\( MC, yo \)) being few enough in number that their automatic identification is not a major issue. For this reason, words of 2 characters or less were not considered in the classification task.

A list of stop words, that is, words that are removed from consideration by the classifier in the evaluation phase, is commonly used in classification tasks to remove cases where the classifier would not be reasonably expected to perform correctly. The list of stop words used in the present research was informed by the analysis of errors that classifiers made on the development set along with words predicted to be problematic. The list of stop words used in the present study is comprised of the following:
The list of wordforms found in the German CELEX training data that were also found in the English CELEX training data; these are primarily internationalisms, words borrowed or formed from Greek, Latin, or French in both English and German usage (*Amateur, Anarchist, Gastritis*), and shorter homographs (e.g. *man*, Ger. 'one [impersonal pronoun]', *hat*, Ger. inflected form of *haben*, 'to have'). (4,156 words)

A list of all 3-character words without vowels. (9,260 words, largely nonsense)

A list of wordforms appearing in the names of hip hop artists. This was compiled from several sources: American artist names were taken from a large list of top albums compiled by a hip hop fan, supplemented by a top 100 list of hip hop songs from the 2000s, which was selected to fill a perceived gap in the first list, which is skewed heavily toward albums from the 1980s and 1990s.23,24 (467 words) German artist and record label names were taken from a Wikipedia list of German hip hop artists.25 (552 words)

The most frequent 100 wordforms from a large German news corpus and the 50 most frequent from the Project Covo corpus. These include almost exclusively short function words.26 (150 words)

With redundancies removed, the stop words file totals 14,518 words (roughly 2/3 of which are the nonsense 3-letter words). In the 10,000 word development set, the implementation of preprocessing reduced the number of tokens under consideration to 4,680 words.

4.2.5 n-gram classification

As discussed above, a 1-gram model (e.g., <dog> would have the feature vector ‘d:1 o:1 g:1’) is a reasonable starting point or baseline against which to compare precision, recall, and F1 score for various classifiers, when training data, development set, and stop words are kept constant. To this end, a variety of different n-gram features were extracted, and different combinations of features were extracted to create classification systems on different feature vectors. As an example, the

23 “The 500 greatest hip hop albums plus the other ones that are honorable mention – Rate Your Music” [http://rateyourmusic.com/list/ChrisPC/the_500_greatest_hip_hop_albums__plus_the_other_ones_that_are_honorable_mention_] Accessed 6 June 2011.
The word *freestylen*, 'to freestyle [rap]' would have the following feature vectors in a 1-, a 2-, and a 3-
gram model:

(1a) *freestylen* in 1-grams: ['f:1', 'r:1', 'e:3', 's:1', 't:1', 'y:1', 'l:1', 'n:1']

(1b) *freestylen* in 2-grams: [#f:1', 're:1', 'ee:1', 'es:1', 'st:1', 'ty:1', 'le:1', 'en:1',
'n#:1']

(1c) *freestylen* in 3-grams: [#fr:1', 'fre:1', 'ree:1', 'ees:1', 'est:1', 'sty:1', 'tyl:1',
'yle:1', 'len:1', 'en#:1']

(1d) *freestylen* in 1- through 3-grams: ['f:1', 'r:1', 'e:3', 's:1', 't:1', 'y:1', 'l:1', 'n:1', '#f:1', 're:1',
'ee:1', 'es:1', 'st:1', 'ty:1', 'le:1', 'en:1', 'n#:1', '#fr:1',
'fre:1', 'ree:1', 'ees:1', 'est:1', 'sty:1', 'tyl:1', 'yle:1',
'len:1', 'en#:1']

As seen in (1a-d), from 2-grams on, hash marks <#> are used as a null character at word
boundaries, allowing the classifier to take word-initial and word-final position into account when
determining probabilities. Classifiers were trained on cumulative sets of features, meaning that
the '1-2-gram classifier' would involve both the 1-gram and 2-gram feature sets, the '1-3-gram
classifier', as in (1d) above, would be trained using the 1- through 3-gram feature sets. The '2-5-
gram classifier', then, would operate based on the extraction of the 2, 3, 4, and 5-gram feature
sets from the data. A comparative overview of the performance of the different combinations of
feature sets on the development set is presented in Section 4.5.2.

4.3 Moving to the next level: addressing the challenges of hybrid data

After evaluation of the initial n-gram classifiers, the highest performer was chosen as a baseline
for further improvement. Keeping in mind the goal of this project, that is, to develop a classifier
which can most reliably identify a maximally complete set of anglicisms from the MZEE corpus,
a number of augmentations to the classifier's performance were devised and implemented at
various points in the classification system. Each of these extensions to the system was assigned a
shorthand name and treated as an independent variable for individual and combined evaluation
of performance on the development set. This evaluation is presented in Section 4.5.3.
4.3.1 Dictionary classification

Given the availability of the CELEX wordlists in German and English, one possible way to improve the performance of a classification system is to build on sets of words already unambiguously identified, i.e., to automatically classify words appearing in either the CELEX English wordlist or the CELEX German wordlist as English or German, respectively, ignoring homographs which appear in both wordlists. As an example of previous implementation, this was essentially the first step in Alex's (2008) classifier. For the sake of simplicity, this was implemented in the present study as a re-classification of the data already handled by the MALLET classifier. While it would be optimal in terms of efficiency to use dictionary classification on the data beforehand—using the MALLET classifier only on those terms not found in the dictionaries—the results are in any case equivalent. In comparing classifiers later in this chapter, the inclusion of dictionary classification in a test condition is indicated by 'DC'.

4.3.2 Morphological analysis: stripping affixes

As mentioned in the introduction of this chapter, I hypothesized that the performance of the classifier could be significantly improved if certain German morphological affixes were ignored in the extraction of features. In the example used earlier, the participial form *geflowt* 'flowed', the English stem *flow* takes the German participial affixes *ge-* and *-t*. Typologically, German is a moderately inflected language, but it shows a higher degree of inflection than English. As a result, English loanwords in the corpus are often morphologically adapted, acquiring German inflectional or derivational morphology. Affix stripping is a process for which a number of algorithms have been developed for various languages. A pre-existing German stemming algorithm (Porter 2001) was considered for this purpose, but the algorithm ultimately had too many drawbacks, not least of which was its failure to deal with prefixes entirely, and it was determined that an original affix stripper could more profitably be tailored to the present research.

The affix stripper was developed as follows: a small list of English stems was compiled, known from preliminary research (the by-hand analysis presented in Chapter 3) to be relatively common in the data and appear in a number of derived and inflected forms. These stems were: *battle, bite,*
chill, deep, diss, download, flow, pop, and release. Some (battle, bite, chill, deep, diss, flow) could be considered hip hop-related terms, two (download, pop) are more widely established, and release is a term associated generally with multiple musical genres. The MZEE corpus was searched for all words containing these stems (dropping the final <e> on <battle> and <release>), and co-occurring prefixes and suffixes were recorded. Because of the possibility of multiple prefixation or suffixation (e.g. rum-ge-battle or deep-er-en), two prefix positions and three suffix positions were used, but the three suffix positions caused problems with overgeneration (too many non-suffixes were accidentally captured) and were ultimately consolidated into a single list of legitimate compound suffixes to address initial problems with overgeneration. Additionally, paradigmatic gaps were filled after consultation with a German linguistic reference (Fagan 2009). The outer prefix list consisted largely of prepositional prefixes (e.g. über-, ab-, zu-), and the inner prefix position had only ge- as a possibility, thus allowing for either prefixed (in the absence of prepositional prefixes) or infixed ge-. All prefixes and suffixes appearing with the short list of stems in the MZEE corpus were recorded, as demonstrated with selected affixes in Table 4.2.27

<table>
<thead>
<tr>
<th>Prefix 2</th>
<th>Prefix 1</th>
<th>STEM</th>
<th>Suffix 1</th>
<th>Suffix 2</th>
<th>Suffix 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>rum-</td>
<td>ge-</td>
<td>-flow-</td>
<td>-e</td>
<td>-e</td>
<td>-e</td>
</tr>
<tr>
<td>ab-</td>
<td>-</td>
<td>-</td>
<td>-t</td>
<td>-er</td>
<td>-er</td>
</tr>
<tr>
<td>zu-</td>
<td>-</td>
<td>-</td>
<td>-st</td>
<td>-es</td>
<td>-es</td>
</tr>
<tr>
<td>pseudo-</td>
<td>-</td>
<td>-</td>
<td>-los</td>
<td>-en</td>
<td>-en</td>
</tr>
<tr>
<td>un-</td>
<td>-</td>
<td>-</td>
<td>-bar</td>
<td>-em</td>
<td>-em</td>
</tr>
<tr>
<td>...</td>
<td>-</td>
<td>-</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 4.2. Selected prefixes and suffixes demonstrate the use of different affix positions in the original affix-stripping schema.

These lists were used in an altered version of the feature extraction process to 'silence' common prefixes and suffixes in the resulting feature vectors, ensuring each time that a stem of at least 4

27 It is important to note here that I do not posit these as the only possible prefix and suffix positions for the German language, but as an approximation of the reasonable common maximum encountered in the majority of declined and inflected forms. Unlike a true stemmer, in which the goal is to produce the correct word stem, the purpose of this affix stripper is solely to improve the performance of the classifier on the MZEE.com data.
28 Translations of individual affixes are not provided here because of the high amount of syncretism in such affixes.
characters remained, so as not to yield an empty feature vector, for, e.g., \textit{get}. An example follows using the wordform \textit{abgechillt} 'chilled out', 'relaxed' and a feature extraction script using only 4-grams:

(2a) \textit{abgechillt} \rightarrow ['#abg:1', 'abge:1', 'bgec:1', 'gech:1', 'echi:1', 'chil:1', 'hill:1', 'illt:1', 'ill#:1']

(2b) \textit{abgechillt} 'relaxed' \rightarrow \textbf{prefix ab- removed} \rightarrow \textit{gechillt} \rightarrow \textbf{suffix -t removed} \rightarrow \textit{gechill} \rightarrow \textbf{prefix ge- removed} \rightarrow \textit{chill} \rightarrow ['#chi:1', 'hill:1', 'ill#:1']

In (2a-b), the result of the feature extraction script would be \textit{abgechillt} in either case—only the feature vector associated with the word would be changed.

An important note is that in contrast to the stemmer presented by Porter (2001), the morphological process of umlaut (the vowel change in an inflected or derived word) was not reversed when removing suffixes that trigger said process. The reasoning behind this decision is that, in contrast to the general purpose of a true stemmer, i.e., to identify the lexical stem from multiple inflected or declined wordforms, the purpose of the present affix stripper is simply to silence German bound morphemes which might appear with an English word. To illustrate why the retention of umlaut might be desirable in this case, consider the following adjectives alongside their superlatives:

(3a) \textit{kalt} 'cold' \rightarrow \textit{kälteste} 'coldest'

(3b) \textit{wack} '(Hip Hop English) bad' \rightarrow \textit{wackste} 'worst'

The removal of the -(e)ste superlative suffix from (3a) will leave \textit{kält}, which, while not a German word, is more readily identifiable to the classifier as such because of the character \textless{ä}\textgreater{}, which does not appear in English words. On the other hand, while \textit{wack} is, phonologically, a possible German word, the removal of -\textit{ste} could reduce the likelihood that this word is misclassified as German, assuming a 4-gram classifier that recognizes \textless{ste#}\textgreater{} as a feature of more German words (where word-final -\textit{ste} is a productive morpheme) than English words (where only a limited set exists, like \textit{taste} and \textit{waste}).

The affix stripper was ultimately implemented as part of the feature extraction script as two mutually-exclusive variables, each evaluated separately and in combination with the other
augmentations to the classification system, but not with each other. The first method (indicated in an experimental condition by 'AS1') was to use the affix stripper only on those words not in the CELEX English or German dictionaries, and then only in the development set—the training data had all features extracted normally. This could be considered a less-invasive application of the affix stripper, as fewer feature vectors were changed.

The second, and more extensive, application of the affix stripper, indicated by the label 'AS2' in an experimental condition, involved the application of the stripper to all words in both the development set and the training data, changing many more feature vectors, but having the advantage of consistency between the classifier's knowledge and its task. While less information was ultimately available to the classifier in this application, this information was always consistent between the training data and the development set.

4.3.3 Hybrid compound analysis: cutting compounds

Nominal and adjectival compounding is a common process in German, and one which is often commented on by non-native speakers. While examples such as (4) abound in humorous portrayals of German, such extreme instances of compounding are generally more indicative of highly official writing, rather than everyday natural language production. (4) has here been separated into lexemes by the addition of vertical bars; the three |s| segments can be considered epenthetic for the purposes of the present discussion.

(4) Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz
    Rind|fleisch|etikettierung|s|überwachung|s|aufgaben|übertragung|s|gesetz
    beef |meat   |labeling       | |oversight        |  |duties     |delegation   |  |law
'The law concerning the delegation of duties for the supervision of the labeling of beef'

While examples like this are rare, the compounding of two adjectives or nouns is relatively common, and several examples found during the by-hand preliminary analysis of the 10-thread subcorpus in Chapter 3 demonstrate the ways in which English borrowings can be enlisted into such a process to create bilingual or hybrid compounds. In (5a-e) I have added vertical bars to separate lexemes:
Because of the hybrid nature of these wordforms, the classifier would be likely in each case to pick up on character sequences like <lich#>, <pf>, and <ö>, which are strongly indicative of a German label. However, for the purposes of the present study, these forms are considered hybrid anglicisms, and as such, their automatic identification is highly desirable.

To properly capture these hybrid anglicisms, a compound-cutting algorithm was devised and integrated into the classification script. The first step is to check the label of the entire word—if it is already labeled E, the compound cutter does not apply. The process splits a word over a certain length \( l \) into two portions in a number of ways, according to a variable \( n \), which indicates the number of cuts in each direction from the center of the word. \( n \), then, is defined as the number of word boundaries from the center that the process creates in each direction. If \( n \) is 2, for example, and the word in question is \( \text{Soundhörer} \), the process works as follows.

(6) \[ \text{Sound} | \text{hörer} \]

Center: \( \text{sound} | \text{hörer} \) ← correct version
Center +1: \( \text{soundhö} | \text{rer} \)
Center +2: \( \text{soundhö} | \text{rer} \)
Center -1: \( \text{soun} | \text{dhörer} \)
Center -2: \( \text{sou} | \text{ndhörer} \)

Another variable, \( m \), dictates the minimum length of each part. With \( m = 4 \), for example, the cutter would stop after cutting the word one character in each direction in (6) above, because the length of the strings <rer> and <sou> would prevent it from proceeding. After cutting the word, the algorithm passes each word part to the classifier. If one of the parts is labeled as E by the classifier, the compound cutter stores that information along with the given confidence, a number between 0 and 1 that the classifier also reports. After this is done with all possible word parts, the final variable comes into play: a target confidence level \( c \). One of three things then happens: one,
none of the word parts were labeled E, in which case the entire compound remains labeled G; two, the maximum confidence for any E-labeled word part fails to exceed the target confidence \( c \), in which case the entire compound remains labeled G; or three, one of the E-labeled word parts exceeds the target confidence \( c \), in which case the entire compound is re-labeled E in the classifier's output.

To improve the performance of the compound cutter, the variables \( l, n, m, \) and \( c \) were optimized through a series of evaluations. The best results were obtained with the following parameters: \( l \), the minimum word length for consideration, at 7 characters; \( n \), the number of cuts in each direction from the center of the word, at 3; \( m \), the minimum length of each word part, at 2 characters; and \( c \), the target confidence level set equal to the original confidence level for the compound's G label less 0.3. As a variable during evaluation, the presence of this optimized compound-cutting algorithm in an experimental condition is indicated by the label 'CC'.

4.4 Overview of the classification system

Figure 4.1, below, provides a graphical representation of the classification system. The various augmentations to the system, including the affix-stripping algorithms, the compound-cutting algorithm, and the dictionary classification process, are connected to the system with dashed lines, indicating their optional nature. The dotted-line arrows emanating from the error analysis step at the bottom indicate gradual improvements to the classification system based on the output and performance on the development set.
Figure 4.1. Overview of the classification system used in this dissertation.
After the various competing classifiers had been thoroughly tested, the optimal classifier was selected with the highest-performing additional processes; this was then used to automatically identify candidate anglicisms in the entire MZEE.com corpus.

4.5 Results

In this section, I begin by presenting the results of the analyses described in the previous section, starting with the results of early trials which led to the decision to discard upper-case information in the later trials. I continue by discussing the findings on the use of different $n$-gram feature vectors, and finally, I consider the effects of the variable classifier extensions individually and in combination. Before closing, I discuss the highest-frequency anglicisms identified by the classifier, which are reproduced in Appendix A, and discuss what sort of broad categories they might be said to fall into.

4.5.1 Effects of case on classification

Because German and English have differing capitalization standards in their written form, German having an additional orthographic rule regarding the capitalization of nouns (nouns are capitalized), it may be expected that preserving the case of words in the training data and development sets would improve the classifier's ability to distinguish German content from English content. As mentioned in section 4.2.2, the informal nature of the target data from the MZEE corpus means that any given word could be expected to be found lowercased (flavor), title-cased (Flavor), uppercased (FLAVOR), or even in 'sticky caps' (fLaVoR, FlavoR, FlAvOr, etc.). While the latter cases are perhaps more rare, the frequency with which words begin sentences (prescriptively title-cased in both languages) and the lax application of orthographic rules in an online youth-cultural environment make for irregular use of capitalization. Early trials with classifiers on a 5,000 word preliminary development set showed negligible gains or outright losses in all measures when capitalization was taken into account in the training data and the development set (i.e., when the training data and development set were not universally lowercased, and alternative title-cased and uppercased forms were added to the training data). Table 4.3 and Figure 4.2, below, show the comparison between mixed-case and lower-case classification. It must be noted that while these trials are comparable with one another, given that
the training data, development set, and stop words list was internally consistent, these should not be compared with the classification results found later; the development set used here was only 5,000 words in length, and a reduced list of stop words consisting only of the intersection of the English and German CELEX wordlists was used.

<table>
<thead>
<tr>
<th></th>
<th>1-4-grams, mixed-case</th>
<th>1-4-grams, lower</th>
<th>1-5-grams, mixed-case</th>
<th>1-5-grams, lower</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>58.7%</td>
<td>70.1%</td>
<td>Accuracy</td>
<td>73.9%</td>
</tr>
<tr>
<td>Precision on E</td>
<td>24.9%</td>
<td>31.3%</td>
<td>Precision on E</td>
<td>34.5%</td>
</tr>
<tr>
<td>Recall on E</td>
<td>95.4%</td>
<td>93.6%</td>
<td>Recall on E</td>
<td>94.9%</td>
</tr>
<tr>
<td>F1 on E</td>
<td>39.5%</td>
<td>46.9%</td>
<td>F1 on E</td>
<td>50.7%</td>
</tr>
</tbody>
</table>

Table 4.3. Comparison of mixed-case and lower-case training/development data in early trials.

Because all measures other than recall suffered in both mixed-case conditions, I made the decision after these trials to discard case. This decision also carried an additional benefit: because the training data did not need to be 'filled out' with additional upper- and title-case wordforms, classifier training ran more quickly.
4.5.2 Assessment of n-gram classification

In this stage of the analysis, six classifiers trained on differing n-gram feature vectors were evaluated in their performance on the development set, the results of which are presented in Table 4.4 and Figure 4.3.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Threshold</th>
<th>Precision on E</th>
<th>Recall on E</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-gram</td>
<td>0.5</td>
<td>25.18%</td>
<td>53.59%</td>
<td>33.77%</td>
<td>67.16%</td>
</tr>
<tr>
<td>1-2-gram</td>
<td>0.5</td>
<td>42.01%</td>
<td>84.25%</td>
<td>56.07%</td>
<td>78.91%</td>
</tr>
<tr>
<td>1-3-gram</td>
<td>0.5</td>
<td>49.65%</td>
<td>89.50%</td>
<td>63.87%</td>
<td>83.83%</td>
</tr>
<tr>
<td>1-4-gram</td>
<td>0.5</td>
<td>59.56%</td>
<td>92.96%</td>
<td>72.60%</td>
<td>88.78%</td>
</tr>
<tr>
<td>1-5-gram</td>
<td>0.5</td>
<td>64.86%</td>
<td>93.78%</td>
<td>76.68%</td>
<td>90.88%</td>
</tr>
<tr>
<td>1-6-gram</td>
<td>0.5</td>
<td>65.45%</td>
<td>93.92%</td>
<td>77.14%</td>
<td>91.10%</td>
</tr>
</tbody>
</table>

Table 4.4. Performance of various n-gram classifiers on the development set, threshold .5.

The addition of longer n-gram features to the feature vectors used in each condition improved all measures of performance, but the most significant gains are seen early on. While I hypothesize that the addition of 7-grams to the feature vectors would increase performance slightly, the
addition of features to a feature vector increases the processing required to create each classifier as well as the complexity of the classifier. In addition, 87.6% of the data consists of tokens of seven characters or less, and thus only a small set of the data would actually be affected by these extra features. The results above, then, demonstrate that the 1-6-gram model is the most reliable of the classifiers tested in this analysis.

At this juncture, I will explain the concept of threshold as it is used here. During this part of the analysis, the method of classifier evaluation was altered. As mentioned in the previous section, the classifier's output consists of a wordform, its label (G or E) and a third piece of data: a confidence measure. This value, which is represented as a decimal number between .5 and 1, indicates the classifier's computation of the probability that its chosen label is accurate. Essentially, this value was recalculated as a value on a scale from 0 to 1 as shown in (7) below:

(7)  

<table>
<thead>
<tr>
<th>Original:</th>
<th>G 1 0.75 0.50 0.75 1 E</th>
</tr>
</thead>
<tbody>
<tr>
<td>New:</td>
<td>G 0 0.25 0.50 0.75 1 E</td>
</tr>
</tbody>
</table>

Under this transformation, the label provided by the classifier is based on a threshold of .5—any word with a new confidence value of less than .5 is labeled G, and any word with a new confidence value greater than .5 is labeled E. By adjusting this threshold and essentially re-labelling data, a different measure of optimal performance can be created. Given that the purpose of this classifier is to reliably identify a high percentage of anglicisms in the corpus (i.e., attaining a high degree of recall on E) while reducing the amount of by-hand correction needed in the output (i.e., increasing precision on E), it was decided that classifiers be compared at a measure of maximum precision at >95% (or in some cases where this was unattainable, >90%) recall on E for the development set. Measures of precision and recall are not entirely independent—Figure 4.4 plots precision and recall (along with accuracy and F1—the harmonic mean of precision and recall) of the 1-6-gram model across 98 values of threshold (along the x-axis):
Figure 4.4. Measures of classifier performance for the 1-6-gram classifier with varying threshold.

As the confidence threshold at which words are labeled G or E changes, so too do the evaluation statistics for the classifier in question. As the threshold approaches 1, more words' confidence values fall below the threshold, and those words are classified as G. Likewise, fewer words have confidence values high enough to be classified as E. The result is that, as the threshold is adjusted upward, recall decreases as more words which should be labeled E are labeled G instead, and precision increases as fewer of the words labeled E are actually German words.

Another way to visualize the covariation of precision and recall is by comparing the two, as in Figure 4.5, which shows the results of precision vs. recall for the baseline 1-6-gram classifier, along with the results of precision vs. recall for the following conditions: AS2 (affix-stripping applied to both training and test data), CC (compound-cutting), and those two parameters in combination:
Figure 4.5. Precision plotted against values of recall on E for the baseline 1-6-gram model, the model with affix-stripping, the model with compound-cutting, and the model with both affix-stripping and compound-cutting.

Here, precision on E for the classifiers is generally shown to decrease as recall increases. At the risk of presaging later results, it should be noted that the ideal classifier has both high precision and high recall—that is, values which, plotted on the graph above, are maximally close to the upper right corner. As seen in Figure 4.5, affix-stripping and compound-cutting both improve the performance of the baseline classifier, but the combination of the two increases performance even more, especially at high recall, which is desirable in this case. Each classifier built using the n-gram feature vectors was evaluated on its performance at thresholds from 0.01—where only those words which the classifier is at least 99% certain are German are labeled G—to 0.99—where only those words which the classifier is at least 99% certain are English are labeled E. Thus, in Table 4.5, below, results are reported for each n-gram model at the threshold level with maximum precision where recall was greater than 95%.
### Table 4.5. Performance of various n-gram classifiers on the development set, maximum precision at >95% recall.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Threshold</th>
<th>Precision on E</th>
<th>Recall on E</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-gram</td>
<td>0.05</td>
<td>16.82%</td>
<td>95.03%</td>
<td>28.58%</td>
<td>24.20%</td>
</tr>
<tr>
<td>1-2-gram</td>
<td>0.14</td>
<td>28.63%</td>
<td>95.58%</td>
<td>44.06%</td>
<td>61.25%</td>
</tr>
<tr>
<td>1-3-gram</td>
<td>0.14</td>
<td>33.71%</td>
<td>95.17%</td>
<td>49.78%</td>
<td>69.34%</td>
</tr>
<tr>
<td>1-4-gram</td>
<td>0.17</td>
<td>43.01%</td>
<td>95.17%</td>
<td>59.24%</td>
<td>79.08%</td>
</tr>
<tr>
<td>1-5-gram</td>
<td>0.27</td>
<td>54.60%</td>
<td>95.03%</td>
<td>69.35%</td>
<td>86.58%</td>
</tr>
<tr>
<td>1-6-gram</td>
<td>0.28</td>
<td>62.15%</td>
<td>95.03%</td>
<td>75.15%</td>
<td>89.95%</td>
</tr>
</tbody>
</table>

As in Figure 4.3, where the threshold was held constant at .5, the precision in Figure 4.6 improves as larger n-gram features are added—but these improvements in precision are more even, as recall is held above 95%. The 1-6-gram classifier, which performed best of the n-gram classifiers tested, was selected as the baseline classifier for further performance improvement.
4.5.3 Effects of classifier extensions

After the 1-6-gram classifier was selected for further improvement, different means of improving the results of this classifier (as outlined in section 4.3) were tested individually and in combination. In this section, results are presented first for the entire development set (4,623 word tokens after the removal of stop words), then for unique word types from the development set (2,738 word types after the removal of stop words), and finally for unique word types which are unknown to the classifier, i.e., those which do not appear in the training data (665 word types after the removal of stop words and words appearing in the training data). Conditions are reported as follows:

- 'Baseline': unaltered results from the 1-6-gram classifier discussed in the previous section.
- 'AS1': affix-stripping applied to words in the development set which are not found in either CELEX wordlist.
- 'AS2': affix-stripping applied to all words in the development set and the training data.
- 'CC': compound-cutting extension implemented in the classifier.
- 'DC': dictionary classification is applied to the classifier's output, re-labelling some words if they appear in the CELEX wordlists.

Conditions involving dictionary classification are reported in separate tables from other conditions for the sake of readability and because dictionary classification is not applied in the 'unknown words' analysis presented later. As the dictionaries in question were part of the training data, words in the development set which appear in the dictionaries are exempted from that analysis—therefore, dictionary classification has no effect.

Tables 4.6 and 4.7, along with Figure 4.7, below, report the performance of various conditions on the entire development set. The primary measure to consider here, and in the following tables, is precision on E, as the results are reported at >95% (or in the final case, >90%) recall.
### Table 4.6. Measures of classifier performance on all word tokens with maximum precision at >95% recall, no dictionary classification.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Threshold</th>
<th><strong>Precision on E</strong></th>
<th>Recall on E</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.28</td>
<td>62.15%</td>
<td>95.03%</td>
<td>74.42%</td>
<td>89.95%</td>
</tr>
<tr>
<td>AS1</td>
<td>0.69</td>
<td>66.47%</td>
<td>95.03%</td>
<td>78.23%</td>
<td>91.54%</td>
</tr>
<tr>
<td>AS2</td>
<td>0.27</td>
<td>59.19%</td>
<td>95.17%</td>
<td>72.99%</td>
<td>88.74%</td>
</tr>
<tr>
<td>CC</td>
<td>0.68</td>
<td>65.15%</td>
<td>95.03%</td>
<td>77.30%</td>
<td>91.07%</td>
</tr>
<tr>
<td>AS1+CC</td>
<td>0.69</td>
<td>64.25%</td>
<td>96.82%</td>
<td>77.25%</td>
<td>90.88%</td>
</tr>
<tr>
<td>AS2+CC</td>
<td>0.37</td>
<td>59.87%</td>
<td>95.44%</td>
<td>73.59%</td>
<td>89.04%</td>
</tr>
</tbody>
</table>

### Table 4.7. Measures of classifier performance on all word tokens with maximum precision at >95% recall, with dictionary classification.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Threshold</th>
<th><strong>Precision on E</strong></th>
<th>Recall on E</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline+DC</td>
<td>0.28</td>
<td>63.94%</td>
<td>95.03%</td>
<td>76.44%</td>
<td>90.63%</td>
</tr>
<tr>
<td>AS1+DC</td>
<td>0.77</td>
<td>66.76%</td>
<td>95.44%</td>
<td>78.57%</td>
<td>91.67%</td>
</tr>
<tr>
<td>AS2+DC</td>
<td>0.6</td>
<td>69.08%</td>
<td>95.03%</td>
<td>80.00%</td>
<td>92.40%</td>
</tr>
<tr>
<td>CC+DC</td>
<td>0.69</td>
<td>67.06%</td>
<td>95.30%</td>
<td>78.72%</td>
<td>91.76%</td>
</tr>
<tr>
<td>AS1+CC+DC</td>
<td>0.76</td>
<td>68.12%</td>
<td>95.03%</td>
<td>79.35%</td>
<td>92.09%</td>
</tr>
<tr>
<td>AS2+CC+DC</td>
<td>0.64</td>
<td>69.81%</td>
<td>95.17%</td>
<td>80.54%</td>
<td>92.64%</td>
</tr>
</tbody>
</table>

In Table 4.6, the highest performers, using the measure of maximum precision at >95% recall, are, in order, AS1, CC, and AS1+CC. While it was expected that the combination of different extensions would yield diminishing returns, it was not expected that the combination of two extensions, as in the case of AS1+CC, would actually yield (slightly) lower performance than either extension operating alone, indicating that the two augmentations may be interfering in their operation. In this case, it is reasonable to suspect that the affix stripper's work (which, ideally, should increase both precision and recall) is being undone by the compound-cutter, which universally increases recall, possibly at the cost of precision. However, in Table 4.7, when dictionary classification is brought into the mix, the highest results of all are achieved.
Figure 4.7. Maximum precision on E at >95% recall, all conditions; all tokens of development set.

The second application of affix-stripping, i.e. to all words in the training data and the development set, seems to work especially synergistically with dictionary classification. Dictionary classification, considered across conditions, improves the measure of precision at >95% recall by an average (mean) of 4.61 percentage points, compared to the more modest mean gain from compound-cutting of 1.21 points, the mean gain from AS1 over non-AS conditions of 1.80 points, or the small mean loss from the addition of AS2 to non-AS conditions of -0.09 points. This latter effect can most probably be attributed to the loss of information to the classifier in AS2 conditions due to the loss of character sequences corresponding to German affixes on all words. However, the highest-performing condition of all, at 69.81% precision on E at >95% recall, was AS2-CC-DC, one of the AS2 conditions.

When wordform types, rather than tokens, are considered, similar results are found, with the AS2 (without DC) conditions making modest gains or losses compared to their non-AS counterparts, and the AS2+DC conditions breaking away from the pack, as seen in Tables 4.8 and 4.9, along with Figure 4.8:
Table 4.8. Measures of classifier performance on all word types with maximum precision at >95% recall, no dictionary classification.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Threshold</th>
<th>Precision on E</th>
<th>Recall on E</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.16</td>
<td>57.66%</td>
<td>95.03%</td>
<td>71.77%</td>
<td>86.52%</td>
</tr>
<tr>
<td>AS1</td>
<td>0.43</td>
<td>61.20%</td>
<td>95.03%</td>
<td>74.45%</td>
<td>88.24%</td>
</tr>
<tr>
<td>AS2</td>
<td>0.24</td>
<td>59.46%</td>
<td>95.03%</td>
<td>73.15%</td>
<td>87.42%</td>
</tr>
<tr>
<td>CC</td>
<td>0.53</td>
<td>61.36%</td>
<td>95.03%</td>
<td>74.57%</td>
<td>88.31%</td>
</tr>
<tr>
<td>AS1+CC</td>
<td>0.69</td>
<td>62.28%</td>
<td>96.07%</td>
<td>75.57%</td>
<td>88.79%</td>
</tr>
<tr>
<td>AS2+CC</td>
<td>0.34</td>
<td>59.23%</td>
<td>95.03%</td>
<td>72.97%</td>
<td>87.30%</td>
</tr>
</tbody>
</table>

The performance gains when the AS2 and DC variables co-occur can be explained thus: because both the development set and the training data for the classifier are stripped of affixes in AS2, unlike in AS1, where only certain words in the development set are stripped of affixes, non-dictionary words benefit from this consistency, which is not present in AS1. At the same time, the loss of information to the classifier hurts overall performance when dictionary classification is not in place. However, the final dictionary classification corrects a number of the errors made, leaving only non-dictionary words to benefit from AS2.

Table 4.9. Measures of classifier performance on all word types with maximum precision at >95% recall, with dictionary classification.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Threshold</th>
<th>Precision on E</th>
<th>Recall on E</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline+DC</td>
<td>0.16</td>
<td>61.53%</td>
<td>95.03%</td>
<td>74.69%</td>
<td>88.38%</td>
</tr>
<tr>
<td>AS1+DC</td>
<td>0.48</td>
<td>62.70%</td>
<td>95.03%</td>
<td>75.56%</td>
<td>88.91%</td>
</tr>
<tr>
<td>AS2+DC</td>
<td>0.29</td>
<td>65.85%</td>
<td>95.03%</td>
<td>77.80%</td>
<td>90.21%</td>
</tr>
<tr>
<td>CC+DC</td>
<td>0.57</td>
<td>63.93%</td>
<td>95.03%</td>
<td>76.44%</td>
<td>89.43%</td>
</tr>
<tr>
<td>AS1+CC+DC</td>
<td>0.69</td>
<td>63.45%</td>
<td>96.69%</td>
<td>76.62%</td>
<td>89.35%</td>
</tr>
<tr>
<td>AS2+CC+DC</td>
<td><strong>0.62</strong></td>
<td><strong>66.81%</strong></td>
<td><strong>95.03%</strong></td>
<td><strong>78.46%</strong></td>
<td><strong>90.58%</strong></td>
</tr>
</tbody>
</table>

The performance gains when the AS2 and DC variables co-occur can be explained thus: because both the development set and the training data for the classifier are stripped of affixes in AS2, unlike in AS1, where only certain words in the development set are stripped of affixes, non-dictionary words benefit from this consistency, which is not present in AS1. At the same time, the loss of information to the classifier hurts overall performance when dictionary classification is not in place. However, the final dictionary classification corrects a number of the errors made, leaving only non-dictionary words to benefit from AS2.
As a final comparison of conditions, I present in Table 4.10 the performance of the classifier augmentations on a set of 1755 ‘out of vocabulary' word types, i.e. wordforms not found in the training data, the stopwords list, or the development set. Note that most common German and English words are thus exempted, which in part accounts for the low recall relative to the results for all wordform types reported above; results are reported at the threshold with optimal results on all types. Finally, dictionary classification has no effect for this set because words in the dictionaries (the CELEX wordlists) are exempted from classification.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Threshold</th>
<th>Precision on E</th>
<th>Recall on E</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.16</td>
<td>31.58%</td>
<td>84.96%</td>
<td>46.04%</td>
<td>66.02%</td>
</tr>
<tr>
<td>AS1</td>
<td>0.48</td>
<td>32.21%</td>
<td>84.96%</td>
<td>46.72%</td>
<td>66.92%</td>
</tr>
<tr>
<td>AS2</td>
<td>0.29</td>
<td>36.23%</td>
<td>84.96%</td>
<td>50.79%</td>
<td>71.88%</td>
</tr>
<tr>
<td>CC</td>
<td>0.57</td>
<td>34.04%</td>
<td>84.96%</td>
<td>48.61%</td>
<td>69.32%</td>
</tr>
<tr>
<td>AS1+CC</td>
<td>0.69</td>
<td>34.88%</td>
<td>92.92%</td>
<td>50.72%</td>
<td>69.17%</td>
</tr>
<tr>
<td>AS2+CC</td>
<td>0.62</td>
<td><strong>36.50%</strong></td>
<td><strong>84.96%</strong></td>
<td><strong>51.06%</strong></td>
<td><strong>72.18%</strong></td>
</tr>
</tbody>
</table>

Table 4.10. Measures of classifier performance on out-of-vocabulary types with maximum precision at >90% recall.
The AS2 conditions again achieve the highest precision on E, confirming that stripping affixes from both the training data and development set works best for those words not found in the CELEX dictionaries (which are part of the training data). The compound-cutting algorithm is also seen to have a small positive effect in all cases.

In summary, all hypothesized improvements to the classifier (broadly speaking) had some positive effect, and, in terms of word types, the most effective combination, AS2-CC-DC, saw more than a 9 percentage point gain in the measure of maximum precision on E at >95% recall over the unaltered 1-6-gram baseline. In addition, this value represents a 50 percentage point gain in maximum precision on E at >95% recall over the more primitive baseline of a 1-gram classifier. Ultimately, this translates to a major reduction in by-hand identification of anglicisms on the MZEE corpus, enabling a much broader range of anglicisms to be identified for later analyses than would otherwise be possible.

4.6 Corpus frequency of anglicisms

To apply the findings of the previous section to the goals of the present research, I decided to run the final classifier on all word types in the MZEE list corpus with the second version of affix-stripping (AS2), compound-cutting (CC), and dictionary classification (DC) in place, at a threshold of .62, the configuration represented in the bold face final rows of Tables 4.9 and 4.10 above. The list of stop words was also exempted from analysis at this point. As this constitutes an application of the classifier, rather than an evaluation, words labeled G or E in the hand-coded development set were also re-classified as such, saving more time in the by-hand analysis. Beginning with the highest-frequency candidate anglicisms (i.e., words labeled E by the classifier), the results then were hand-corrected.

The final classifier system identified 94,066 candidate anglicism wordform types; of these, 56,109 (59.6%) were *hapax legomena*, occurring only once in the corpus. Another 12,624 (13.4%) were *dis legomena*, occurring only twice. While hand-correcting the entire anglicism list was not feasible, the identification of anglicisms occurring above a certain frequency threshold was performed. 1,415 anglicism candidates occurred with a corpus frequency of at least 100, that is, a relative frequency of roughly 8 per million tokens in the corpus. Candidate anglicisms with
absolute frequency greater than 100 account for 63.2% of all candidate anglicism tokens. Overall, types with absolute frequency greater than 100 comprise 90.5% of the corpus data.

During hand-correction, it was found that 850 (60.1%) of the candidates considered met the criteria for inclusion in further study; another 442 (31.2%) were either established anglicisms, place names, artist names, or other foreign material, i.e., wordforms which the classifier unsurprisingly identified as English-derived; a final 123 (8.7%) were misclassified German words, most of which were non-standard or 'eye-dialect' orthographic forms like <nem> for <em>anglicism tokens with frequency over 100 were retained, we can estimate that if 68.0% of all the candidate anglicism tokens are novel anglicisms, then there are roughly 715,625 anglicism tokens in the corpus—5.7% of tokens in the corpus are anglicisms. This is a few percentage points lower than the percentage of anglicism tokens found in the by-hand analysis in Chapter 3, which was 7.4%.

It is possible that the 10-thread subcorpus used in that analysis had a higher rate of anglicism use than would be found for other portions of the corpus; on the other hand, the classifier was set for only 95% recall on the development set (i.e., missing 5% of anglicism tokens in the data). Additionally, the methods used here, which exempted a long list of stop words not used in the by-hand analysis, could be affecting these estimates. If the stopwords and other words with a character length less than three are removed from the corpus, we find that the identified anglicisms represent 10.0% of remaining tokens with absolute frequency greater than 100. Finally, anglicism frequency in the hand-coded development set provides another estimate: 8.2% of tokens in the development set were coded as anglicisms. Taking both the earlier by-hand analysis and the varying estimates yielded by classification into account, I would suggest that a confident estimate of (novel) anglicism frequency in the corpus is 4-10%, with the 6%-8% range
being most likely to be accurate. The hand-corrected list of 850 anglicisms from the classifier output is reproduced in full as Appendix A, and particularly interesting forms from this list will be analyzed in depth in the next chapter.

4.7 Conclusions

To address the primary research question from the introduction to this chapter, automatic classification, which has been shown to be useful in previous literature for standard language identification tasks with more standardized data such as newspaper texts, is shown here to be useful for the identification of anglicisms in less standard texts, and with careful affordances, even variably compounded and inflected forms can be reliably identified. While a system with entirely unsupervised output was not found to be useful for the more precise identification needed here, the classifier successfully weeded a large number of non-anglicism forms from the output, which made it possible to identify a large number of anglicisms above a base frequency, a task which would have been prohibitively time-consuming without the classifier described in this chapter.

The classifier presented here was built to be robust in the narrow sense, that is, to deal with noisy natural language data from a specific source, but it was not built to be robust in the broad sense, i.e., to deal with varying data sources, or to function as a general German-English classifier. While the specific classifier presented here is not meant for general purposes, the issues of classifier design addressed here, as well as the techniques used in the present research to improve classifier performance, in particular the customized affix-stripping and compound-cutting algorithms, could be used in the design of future diverse but similarly purpose-built multilingual classifiers.
CHAPTER 5
LOANWORD ADAPTATION IN THE GERMAN HIP HOP COMMUNITY

Dein Gefronte hilft dir nur, die Wahrheit nicht zu peilen, step ans Mikro, doch ich töte deine Crew mit einer Zeile

[...]

Du vergeudest deine Zeit und bitest, ich zerfleische Rapper.
Du und Deichkind, ihr habt was gemeinsam und zwar Scheiße.
Du reimest gut, doch du bist nicht mehr als ein Mittagessen.
Deine Crew behauptet, ich bin nur MC aus Frust, doch ich geh’ ans Mic und rolle über Hip Hop wie ein Bus.

Your fronting only helps you ignore the truth, step to the mic, and I'll kill your crew with a single line

[...]

You're wasting your time and biting, while I'm mauling rappers.
You and Deichkind have something in common, namely, shit.
S-A-V, I split crews in half, liberate fresh flows
You rhyme well but you're nothing more than a lunch.
Your crew claims I'm only an MC out of frustration, but I go on the mic and roll over hip hop like a bus.

–Kool Savas on 'Fehdehandschuh' by Creutzfeld & Jakob, 2000

5.1 Introduction

As part of what Androutsopoulos (2006: 420) describes as a “first wave” of research on computer-mediated communication, much was made of the “hybrid combination of written and spoken features” characteristic of online linguistic forms—this type of research being represented in, e.g., Crystal (2001). More recent analyses (Herring 2004; Androutsopoulos 2006) have taken a critical view of these early investigations, calling for a nuanced and situated approach to language use online, and the result of these efforts is visible in the last decade of scholarship in the paradigm of computer-mediated discourse analysis (CMDA), a program of research laid out in Herring (2004) and exemplified in Androutsopoulos & Beißwenger (2008). While these discourse-centered approaches are of central importance in constructing a more complete and accurate picture of online communication, I suggest in the present chapter that the circumspect study of lexical items and their linguistic features (construed broadly here as inclusive of orthography alongside phonology and morphology) in online environments should not be discarded entirely. In examining English borrowings in German, the linguistic origins of their various morphological, phonological, and orthographic features are the primary source of
popular perceptions of such loans as *fremd*, or 'foreign', and for this reason, the study of lexical items and their linguistic features exists at the crossroads of the social and the linguistic.

The situatedness of English material in hip hop fans' and artists' discourse practices, along with the study of language attitudes and ideologies toward English borrowings, forms a large part of my analysis in Chapter 7; the present chapter instead largely addresses the *what* and the *how* of borrowing, focusing primarily on the interaction of linguistic features from English and German in the adaptation of loanwords in the MZEE corpus, which is described in Chapter 3.

Referencing word types from the frequency-ranked list of anglicisms discovered by the classifier outlined in Chapter 4, I investigate here the linguistic and distributional properties of individual wordforms in the corpus. I begin with the interaction between phonotactics and orthographic variation in the borrowed verb *battlen/batteln* 'to battle', continue by investigating a known case of stylization in the substitution of the grapheme <z> for <s> in plurals, and then investigate <z> along with German morphological additions in orthographic forms of the leave-taking formula *peace*; finally, I explore the range of borrowed verbs ending in *-ed*, finding that the orthographic form <ed> of this English participial morpheme has been extended by German hip hop fans to simple present and certain imperative uses.29 This is an exceptional orthographic extension that, crucially, relies on the syncretism in the suffix *-t* in German and the productive phonological rule of word-final fortition or devoicing in that language. The primary thrust of this chapter, supported by these examples, is that these borrowings, which often appear to be inept or anomalous uses of English, are largely conditioned by German morphophonological rules—resulting in novel, but unexpected, orthographic forms. In light of these findings, I argue here for the importance of attention to orthography in the study of contact-induced language variation and change, particularly in online environments.

5.1.1 Motivation

Sociolinguists might see German as a healthy language with a relatively large speaker population and recognize linguistic change—including contact-induced change—as a linguistic fact rather

29 I remind the reader here of the convention of italicizing linguistic material in a German-language context for purposes of general discussion. Angle brackets <> indicate orthographic forms divorced of other levels of representation. Where relevant, phonological and phonetic forms are indicated by the conventional forward-slashes // and square brackets [], respectively.
than as a threat to the language. However, English borrowings are a major and immediate concern for at least a segment of the German-speaking population, as mentioned briefly in Chapter 1, and as will be discussed in detail in Chapter 7. The analysis in the present chapter is motivated by several of the overall research questions raised in this dissertation: what forms do English-to-German borrowings take in the hip hop community, i.e. how are these borrowings nativized or adapted? Further, what factors affect the borrowability, or the probability of borrowing and continued use, of lexical items? In a descriptive discussion of established English loans in German, Eisenberg (2004: 126) writes:

> Again it should be stressed that we do not want to question in any way the research on language contact which is trying to understand the circumstances under which linguistic forms or meanings are taken over by a language. However it is anything but compelling to talk about the influence of one language on another language and about the consequences of this influence if one does not even know what exactly this influence consists of. [emphasis mine]

The forms under consideration here stand not only to enrich our understanding of the borrowing process, but can also address the extent to which English borrowings are (or are not) integrated into the German language, speaking to the question of language decline or decay in the German political sphere.

5.2 The adaptation of borrowed forms

In terms of the discussion of loanword classification from Chapter 2, the forms discussed in the present chapter are generally of the type which I consider to be visible anglicisms, or to use Onysko's (2007) terminology, those which are formally salient—in every case excluding calques/loan translations and semantic extensions. The specific forms, while not strictly the most frequent anglicisms in the classifier-generated and corrected list (included in this dissertation as Appendix A), are the most frequent anglicisms to exemplify a series of related factors in the adaptation of borrowings which involve the interaction in some way of orthography with morphophonology, an interaction which has been discussed only rarely in previous literature. Androutsopoulos' (2000) examination of punk fanzines, for example, while focused on stylization through orthography, touches on cross-linguistic influence but only in passing.30 The

30 'Fanzines' or simply 'zines' are underground, self-published amateur fan magazines, intended largely for a niche, in-community audience.
borrowings discussed in the present chapter are assessed largely in terms of their integration into German—the addition of (the orthographic representation of) inflectional or derivational morphology being of particular interest.

Aside from the visibility of loanwords, loanwords have been categorized with regard to the general, cross-linguistic borrowability of various word classes or parts of speech, Tadmor et al. (2010) provide a summary of their findings, namely that along a scale of borrowability, nouns are more borrowable than verbs and adjectives, and function words are less borrowable than the former three classes. These findings seem to hold true in terms of the diversity of anglicisms found in Appendix A, although it is worth noting that a number of English function words are also found in the high-frequency anglicism list. While Tadmor et al.’s focus on semantic fields with respect to borrowability is largely tangential to the present project, they find that “not all word meanings are equally often borrowed. Cultural items, such as words relating to religion, clothing, the house, and law, tend to be borrowed often […]” (Tadmor et al. 2010: 232).31 The globalizing culture of hip hop, with its attendant artifacts and practices, then, should be a fertile ground for linguistic borrowing. The research project of Tadmor et al., however, does not include hybridized, nativized, or adapted borrowed forms, which are the primary types of borrowings discussed in this chapter.

While the question of borrowability does intersect with the consideration of a loanword's part of speech, this is problematic in the present analysis. Many of the wordforms in Appendix A fulfill multiple part of speech functions. While some forms, like the participial form gerappt, are monofunctional32 with regard to their part of speech, many, like battle, can be used as a noun (Post #232564, Obwohl ich die Battles net so spannend finde, ‘While I don't find the battles so interesting…’) or as a verb—in this case, a first-person singular present verb (Post #282562, Ich battle dich!!!, ‘I battle you!!!’).33 The common orthographic practice of dropping the final <e> in

31 'Cultural items' is used by the authors largely without specific definition; Tadmor et al. refer here to Swadesh's notion of 'cultural vocabulary' which is defined in opposition to 'basic vocabulary', understood to include “… universal and simple things, qualities, and activities, which depend to the least degree possible on the particular environment and cultural state of the group” (Swadesh 2006 [1971]: 275).
32 It should be made clear at this point that these wordforms are not usually ambiguous in context, but only when taken out of context in, e.g., a corpus analysis.
33 Post numbers given here and throughout this chapter refer to individually numbered posts in the MZEE corpus, from which these examples are drawn.
first-person singular present tense, mirroring the spoken apocope of final schwa means that this potential noun-verb ambiguity affects all verbs with a nominal form identical to the verb stem: (Post #324265, Ich diss euch alle rather than Ich disse euch alle, 'I dis[respect] you all'). The wordform down is similarly multifunctional; it could be a preposition or (separable) verb particle, as in the borrowed participial verb in Post #058058, downgeloadet, 'downloaded', or instead an adjective, as in Post #326002, niemand mit dem wirklich down sein will, 'nobody really wants to be down with him'. Because of this widespread and pervasive multifunctionality, which would necessitate extensive by-hand tagging for accurate part-of-speech identification, I do not quantitatively analyze the frequency with which individual part-of-speech classes in the MZEE corpus are borrowed. I note instead that all four major classes discussed by Tadmor et al. (2010), including the less-borrowable function words, are well-represented in the MZEE corpus as seen in the inclusion of representative forms monofunctional as to part of speech in Appendix A, e.g., the nouns mixtape and hater, the verbs rappt and gedisst, the adjective nice and old, and the function words your and only. The diversity of parts of speech represented in Appendix A, then, indicates the degree of penetration that the practice of borrowing from English has achieved in the German hip hop community.

Returning to the issue of nativization, a number of authors have remarked on the variable morphophonological integration or adaptation of loanwords (Haugen 1950, Poplack et al. 1988, Hock 1991), but only rarely has this adaptation been investigated in-depth and exhaustively categorized for a specific language pairing. Eisenberg's (2004: 128) observational analysis of English borrowings in German is in this case the sole exception. Eisenberg suggests that adjectives, for example, can be categorized along a scale with regard to their inflectional behavior, which he finds “rather heterogeneous, yet anything but chaotic.” Eisenberg also finds that borrowed English verbs can be similarly categorized, and that all readily integrate with German syntax and morphology in a structured and orderly fashion. The anglicisms in Appendix A largely follow German paradigms as regards orthographic gemination (n. Rap, v. rappen) and conjugation (third-person singular present rappt, participial form gerappt) in verbal forms. However, nominal forms like Rap, as Eisenberg (2004: 131) notes for the similar Pin (v. pinnen) and Job (v. jobben), are orthographically aberrant for German, as “such a spelling difference is not permitted in the core vocabulary; it marks a stem as alien”. The reason for this is that the
orthographic gemination would, in a native word, hold for the nominal form as well (i.e., *Rapp, *Pinn, *Jobb). No such issue is found, however, with the form n. Diss v. dissen ('dis[respect]'), which is variably spelled with one or two <s> graphemes in English, but consistently with two <s> graphemes in the MZEE corpus. Unlike Rap, usually pronounced in German as if it were spelled *Räp or *Rep, Diss and dissen fit German orthographic as well as phonological norms (the <i> represents a German /i/), and thus can be considered a less salient form of borrowing, without being composed of German morphemes. With regard to previous research in terms of nativization paradigms, a fairly straightforward integration seems to be the norm for the most common anglicisms in Appendix A. In the following sections, I examine cases of (generally, more salient/visible) anglicisms where the simple application of German morphophonology becomes more complex.

5.3 Phonotactic conditioning of orthographic variation: the case of battle

In this section, I examine an alternation between the orthographic wordforms battlen and batteln (and their variants) as well as a concomitant variation in the noun Battle/Battel, positing that although Battel is far less common than Battle in the MZEE corpus, these forms (along with the verbs) are essentially in a stable complementary distribution. To begin by explaining the orthographic alternation in the verb form, German infinitives are regularly created by the addition of -en to the verb stem. However, when the verb stem ends in /r/ or /l/, the e is deleted, and the infinitive suffix simply becomes -n, as in wandel-n, 'to change'. An explanation of this process is found in Eisenberg (2004: 131). Taking an English word like battle and treating it as a German verb stem creates an orthographic problem: native German verb stems do not typically end in <e>, and the English word battle phonologically ends in /l/, making it an easy target for phonological borrowing with -n. 34 Orthographically, however, there is pressure to change analogically to match German infinitives, many of which end in -eln. This causes a transposition of the <e> and the <l>. However, the new orthographic form batteln conforms to the German paradigm at the price of disrupting the original English orthographic stem, and it is these competing factors that lead to the orthographic variation mentioned above. This is not unique to <battlen/batteln>, and in fact several (non-hip hop-specific) verbs, like <recyclen/recyceln> 'to

34 Observations from ethnographic interviews confirm that, in any case, the spoken version is /batlən/ rather than /batlan/.
recycle' and <googlen/googeln> 'to google' exhibit a similar variation in orthography. For
<recyclen/recyceln>, at least, the latter, more Gemanized form seems to be preferred from a
prescriptive standpoint, but both are considered correct. 35 Which form of battlen/batteln, then,
wins out in the MZEE corpus? Consider the frequencies in Table 5.1:

<table>
<thead>
<tr>
<th>Wordform</th>
<th>Freq.</th>
<th>Wordform</th>
<th>Freq.</th>
<th>Ratio -le- to -el-</th>
</tr>
</thead>
<tbody>
<tr>
<td>battlen</td>
<td>593</td>
<td>batteln</td>
<td>157</td>
<td>3.77/1</td>
</tr>
<tr>
<td>gebattlet</td>
<td>85</td>
<td>gebattelt</td>
<td>24</td>
<td>3.54/1</td>
</tr>
<tr>
<td>battelt</td>
<td>86</td>
<td>battelt</td>
<td>26</td>
<td>3.30/1</td>
</tr>
<tr>
<td>Battle</td>
<td>3299</td>
<td>Battel</td>
<td>80</td>
<td>41.24/1</td>
</tr>
<tr>
<td>Battles</td>
<td>556</td>
<td>Battels</td>
<td>23</td>
<td>24.17/1</td>
</tr>
</tbody>
</table>

Table 5.1. Number of occurrences of battle(n) and battel(n) with related wordforms in the MZEE corpus.

The infinitive forms on the first row (battlen/batteln) are necessarily verbs, as are the participial
forms gebattlet/gebattelt and the forms with verbal morphology battlet/battelt in the next two
rows. In all three cases, the form keeping the English stem intact is preferred by a ratio of about
3.5/1, demonstrating a fairly consistent preference across verb forms. In Figure 5.1, below, the
relative monthly frequency of the infinitive forms is displayed:

---
May 2012; ‘Richtiges recyclen/recyclen im Deutschen | Neue Deutsche Rechtschreibung’ [http://www.neue-
To aid in the correct interpretation of this and the following figures in this chapter, I mention the following details. Each vertical bar represents an adjusted relative frequency of the wordform in question for each calendar month for the period April 2000 through March 2011, and if two or more words are compared, those with the generally greater frequencies are backgrounded. The $x$-axis is labeled with January of each year, and the values on the $y$-axis have been normalized, that is, they represent the adjusted relative frequency as determined by the number of occurrences of a word per million words for a given month (i.e., the word's frequency for that month times one million divided by the corpus size for that month). In addition to the foregrounded/backgrounded bars, Bézier curves are provided to suggest the shape of general trends; as discussed at length in Altmann et al. (2011), content words in large corpora have a tendency to be bursty, i.e., a smooth month-to-month distribution is not to be expected. In this case, a Bézier curve can provide a general sense of the longer-term trends in the data.

Returning to Figure 5.1, it is evident that battle and battle are in a relatively stable competing distribution over time; the slope of the curves is generally similar except for a (very) early spike.
in \textit{batteln} usage. Overall, \textit{battlen} is used more consistently and more often, as the raw frequencies suggest. Both seem to decrease overall over the timespan of the corpus, which is likely a result of the relative decline of the genre of the rap battle itself, a live contest of extemporaneous lyrical skill, as an element of hip hop culture; as hip hop gained mainstream acceptance, the focus shifted to highly-produced and polished recordings, paralleling the gradual shift in several genres of rock away from the early garage-band days. What is perhaps most surprising in Figure 5.1 is that, in over a decade, both forms have survived at a stable ratio, suggesting that neither form lost significant ground to the other in usage preference. This suggests that, even though overall discussion of the concept has declined over time, \textit{battlen} and \textit{batteln} are both well-established and exist in complementary distribution.

As for the final two rows in Table 5.1, comparing \textit{Battle/Battel} and \textit{Battles/Battels}, it is evident that the \textit{-le} orthography is much more strongly preferred in these nominal forms than in the verbal forms above them. In fact, it might be suspected that the wordforms \textit{Battel} and \textit{Battels} are not representative of genuine variation but rather typographical error. However, a comparison of several words of similar absolute frequency and length, and hypothesized typographical errors, reveals that such a suspicion is unfounded:

<table>
<thead>
<tr>
<th>Wordform</th>
<th>Frequency</th>
<th>Wordform</th>
<th>Frequency</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Battle</td>
<td>3299</td>
<td>Battel</td>
<td>80</td>
<td>41/1</td>
</tr>
<tr>
<td>guter, 'good'</td>
<td>3278</td>
<td>*gutre</td>
<td>0</td>
<td>Undef.</td>
</tr>
<tr>
<td>nochmal, 'again'</td>
<td>3273</td>
<td>*nochmla</td>
<td>3</td>
<td>1091/1</td>
</tr>
<tr>
<td>anderes, 'other'</td>
<td>3376</td>
<td>*anderse</td>
<td>1</td>
<td>3376/1</td>
</tr>
<tr>
<td>alten, 'old'</td>
<td>3373</td>
<td>*althe</td>
<td>0</td>
<td>Undef.</td>
</tr>
</tbody>
</table>

Table 5.2. \textit{Comparison of Battle/Battel and wordforms of similar frequency}

To preempt further speculation that this hypothetical typographical error might be specific to the word-final sequence \textit{-le}, or specific to English borrowings, or specific to nouns, consider Table 5.3:
<table>
<thead>
<tr>
<th>Wordform</th>
<th>Frequency</th>
<th>Wordform</th>
<th>Frequency</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Battle</td>
<td>3299</td>
<td>Battel</td>
<td>80</td>
<td>41/1</td>
</tr>
<tr>
<td>viele, 'many'</td>
<td>13529</td>
<td>*vieel</td>
<td>4</td>
<td>3382/1</td>
</tr>
<tr>
<td>coole, 'cool' (f./pl.)</td>
<td>1362</td>
<td>*cooel</td>
<td>1</td>
<td>1362/1</td>
</tr>
<tr>
<td>Style</td>
<td>5219</td>
<td>*Styel</td>
<td>1</td>
<td>5219/1</td>
</tr>
</tbody>
</table>

Table 5.3. Comparison of Battle/Battel and other words ending in -le.

The factor that unites all of the wordforms in the third column of the previous two tables, excepting Battel, is that none of them, when pronounced as hypothetical German words, satisfy the phonotactic constraints of German. Compare these with the doublet Single (2108 hits) / Singel (101 hits), ratio 20.87/1. This similar variation has a ratio roughly half as large as that of Battle/Battel. However, there is one other factor to be taken into account. The wordform battle might not be used solely as a noun, but also as a first-person singular present verb form of battlen, as discussed in section 5.2. As this factor may be expected to skew the ratio, Single/Singel might be more profitably compared with the unambiguously nominal plural forms Battles (556 hits) / Battels (23 hits), ratio 24.17/1. The contrast outlined here between the relative frequencies (< 50/1) of Battle/Battel, Battles/Battels, and Single/Singel, and the hypothesized typographical errors with ratios of > 1000/1 demonstrates that Battel and Singel are in fact low-relative-frequency variations of the more common Battle and Single. An example of Battel used this way in context is given in (1):

(1) [Post #003370]
    Die Frage bezieht sich auf ein sehr bekanntes Freestyle-Battel in Hamburg! ‘The question hinges on a very well-known freestyle battle in Hamburg!’

In Garley (2010), I suggested that the nominal Battel was a result of analogical extension from the more frequent verb form batteln; the frequency of Singel, however, which has no corresponding borrowed verb (there are no occurrences of singeln/singlen in the MZEE corpus) suggests that this earlier analysis was mistaken. Based on the frequency and distributional facts presented here, I suggest instead that orthographic Battel and Singel simply represent nativizations of Battle and Single conditioned by these loanwords' phonological realization in German, and that these forms, though relatively infrequent, can and do survive in stable, complementary distribution with their more common alternatives.
5.4 Orthographic stylization in the MZEE corpus

While the orthographic variation presented in the previous section has an ultimately phonological explanation, the relation of orthography to phonotactics is not as straightforward in every case. The substitution of \(<z>\) for \(<s>\), primarily seen in plural wordforms, has been considered a hallmark of hip hop orthography, as mentioned by Paolillo (2001: 190) in his study of an Indian diaspora chatroom. While Paolillo identifies the use of \(<z>\) as hip-hop related, he notes that it “does not necessarily invoke rap in a direct way” and associates the form with a hacker subculture as well, eventually relating the form to a more general vernacular orthography. This more general view of \(<z>\) fits with Androutsopoulos’ (2000) discussion of *grapheme substitution* as a method of stylization in German punk ‘fanzines’, a genre rich with alternative orthography. The substitution of \(<z>\) for \(<s>\) in English could in many cases be considered a quasi-phonetic spelling, as in the example of \(<\text{dogz}>\) for *dogs*, pronounced [dɔgz], but not in others, as in \(<\text{catz}>\) for *cats*, pronounced [kæts]. In German, however, the grapheme \(<z>\) corresponds to the phoneme /ts/, so the pronunciation of \(<z>\) as /z/, if such a pronunciation is used, is, as Androutsopoulos (2000: 527) notes, “imported from English”. Consider the following three pairs of wordforms, drawn from Appendix A:

<table>
<thead>
<tr>
<th>Wordform</th>
<th>Freq.</th>
<th>Wordform</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beats</td>
<td>9654</td>
<td>Beatz</td>
<td>1826</td>
</tr>
<tr>
<td>Tracks</td>
<td>11231</td>
<td>Trackz</td>
<td>285</td>
</tr>
<tr>
<td>Skills</td>
<td>843</td>
<td>Skillz</td>
<td>856</td>
</tr>
</tbody>
</table>

*Table 5.4. Comparison of plurals with \(<s>\) and \(<z>\) in the MZEE corpus*

Only in the third case, *Skillz*, would the voiced liquid /l/ cause the plural morpheme to become phonetically voiced in English. In the other two cases, *Beatz* and *Trackz*, the orthographic \(<z>\) is essentially spurious in terms of the grapheme’s phonological value in English. Pronouncing the \(<z>\) as [ts], its German value, would be a plausible explanation in the case of *Beatz*, but in the

---

36 *Beats* is not found in Appendix A because, appearing in both the German and English CELEX wordlists, it was removed from automatic analysis. The unusually high frequency of *Skillz* can be traced to an underground hip hop artist who uses *Skillz* as a moniker.
case of *Trackz*, the <z> must necessarily be a stylistic flourish without phonetic value. As in the case of Androutsopoulos' (2000: 527) punk zines,

The crucial motivation for these spelling variants is not phonetic representation, but their indexical or symbolic value as cues of subcultural positioning. In other words, they act as an instruction to interpret the discourse as 'subculturally engaged' or 'hip'.

Likewise, in her study of American hip hop forum discourse, Beers-Fägersten (2006: 40) writes: “final -z is also used to substitute for the inflectional morpheme /s/, even when the phonological environment would not cause voicing” [emphasis mine], indicating that the phenomenon is not unique to German-speaking hip hop fans. There is, however, an additional factor which may make this phenomenon especially prominent in the German case. Word-final devoicing or fortition, a productive phonological rule in German stating that voiced stops and fricatives become voiceless word-finally (Fagan 2009: 23), may be conspiring with the English value /z/ of the grapheme <z> to make the choice of word-final orthographic <z> or <s> moot in terms of pronunciation for German speakers. While this explanation remains provisional at this point, it does suggest the interaction of an orthographic-phonological correspondence from English—that is, providing the English reading of orthographic <z> as phonological /z/ rather than the German reading of <z> as /ts/—with a German phonological rule, namely word-final devoicing, as a result of which the underlying /z/ is realized as phonetic [s]. In this way, then, phonetic realization could play a role in this spelling variation, with the subcultural positioning afforded by the orthographic choice of <z> as a sort of bonus. The next case to be examined will revisit the use of <z>, providing more evidence of the unexpected role that word-final devoicing may play in loanword orthography.

5.5 Orthographic stylization with morphological additions

While it is likely that the practice has its origins before hip hop culture's genesis, the use of *peace* as a formulaic leave-taking can have “specific hip-hop semantics” (Beers-Fägersten 2006: 28). Smitherman (2006: 36) characterizes its use as “Greeting or farewell; originally to indicate uplift, self-love, Black social consciousness.” In the MZEE forum, this leave-taking is found in a large array of forms, as Appendix A attests: *peace* (7855 hits), *peaze* (799 hits), *peaz* (321 hits), *peazen* (317 hits), as well as *peacen*, *peece*, *peeze*, *peez*, *peezn*, *peezee*, *peesen*, *peasze*, *piezen*, *piis*, *piiz*, *piizn*, and a staggering array of variant forms where one or more characters in the above choices
is repeated, as in pеееееееееееееееееееееееенnnn. In addition, idiosyncratic forms like peacigen, peacesen, peaciano, and peacenskofsky are also attested in the MZEE corpus. It should at this point be evident that this single-word leave-taking has been used as something of a creative canvas by the users of the MZEE forum; this sort of stylization is directly connected to the discussion in the previous subsection—and I suggest here that the (sometimes) word-final <z> fulfills the same indexing function described previously. Users' knowledge of both German and English phonological-orthographic mapping is also evident—taking the base form as phonologically /piːs/, the forms with orthographic <ee> display knowledge of a common English representation of /iː/; the forms with orthographic <ii> playfully combine this vowel duplication with the German grapheme for /i/, which is <i>, while avoiding the interpretation of <i> as /ɪ/, which <i> would normally represent in a monosyllabic word like mit, 'with'.

The orthographic <z> reappears in the relatively common alternate form peaze and others like peazen and peaz; however, it is not always orthographically word-final, and could theoretically be pronounced [z] or rather [ts] if the final <e> in peaze is pronounced as it would be in Standard German. Examining a certain variation with the English orthographic convention of the 'silent e', however, illuminates the issue further. To satisfy German orthographic-phonological mapping norms while keeping the loan phonologically intact, the final <e> is dropped in certain orthographic cases, and all else being equal, one would expect this to occur at the same rate regardless of whether the word is spelled with a <c> or with a <z>. The ratios of forms with <e> to forms without <e>, however, are wildly different for peace and peaze, as seen in Table 5.5.

<table>
<thead>
<tr>
<th>Wordform</th>
<th>Frequency</th>
<th>Wordform</th>
<th>Frequency</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>peace</td>
<td>7855</td>
<td>peac</td>
<td>12</td>
<td>654.58/1</td>
</tr>
<tr>
<td>peaze</td>
<td>799</td>
<td>peaz</td>
<td>321</td>
<td>2.49/1</td>
</tr>
</tbody>
</table>

*Table 5.5. Ratios of peace to peac and peaze to peaz in the MZEE corpus.*

To re-iterate, if <z> is a purely stylistic variation in the word peace, with little or no relation to phonology, we should expect the above ratios to be fairly comparable. They are not, and part of the explanation may again lie in word-final devoicing. Taking the English value [z] for <z> (and, for the moment, ignoring the issue of the vowels) peaz could be pronounced, after the application
of word-final devoicing, [pi:s]. *Peac*, on the other hand, is orthographically nonsensical: <c> alone is not a native German grapheme; it generally occurs only in combination: <ch> for /ç/ or /ʃ/; <sch> for /ʃ/. It is also not a native grapheme in word-final position in English, so it is, in this mixed system, doubly aberrant.

A common addition to the base form in the above examples warrants further discussion, and provides a further clue to the phonological value of <z>. Forms like *peacen, peazen, piizen, peesn*, and so forth seem to involve the addition of a syllabic (*-e*)n to the end of the word. This quasi-suffixation is unlikely to be an infinitive verb ending, as *to peace* is semantically problematic. However, an exchange from the forum sheds light on the morphological status of this suffix:

(2a)  [Post #282582]
Wer hilft mir, dass das wort “josen” (bzw. yosen) eingeführt wird?

‘Who’ll help me introduce the word ’josen’ (or rather, yosen)?’

(2b)  [Post #282583]
Das gibts doch bestimmt schon irgendwo wird ja bei vielen wörtern die endung rangehängt wie bei bis densen, peazen, moinsen usw.

‘That’s got to already be out there somewhere, this ending is hung on to many words like bis densen, peazen, moinsen, etc.’

In (2a-b), an MZEE.com user wishes to use introduce the word *josen* or *yosen*, which would be the combination of the hip hop *yo* (or the colloquial German *jo*, ’yes’) with the nonsense suffix – *sen*. The first response is from another user who notes that this -*sen* suffix is already applied to several greetings and leave-takings like *bis dann* ’until then’ > *bis densen, moin* [regionally] ’[good] morning!’ or ’hello!’ > *moinsen*, and, crucially, *peace > peazen*. The inclusion of *peazen* in parallel with the other examples in this post suggests that *peazen* is morphophonologically /pi:s/ + /sen/, further supporting the claim that orthographic *peaz* is phonologically interpreted as /pi:s/ rather than /pi:z/ or /pi:ts/ or another alternative.

In Figure 5.2 below we can see that *peace* is the overwhelmingly preferred form, dwarfing the frequency of the others on a monthly basis. However, it is also evident that most of the
alternative forms exist throughout the time span of the corpus, albeit at much lower rates, suggesting that these are stable alternative forms.

Figure 5.2. Comparison of relative monthly frequency of 4 variants of peace.

The sole exception to this is the form *peazen*, which is compared on a monthly basis with the alternative *peaze* in Figure 5.3, below. Unlike the other alternative versions, *peazen* seems to spike twice in the 2002-2004 time window, and is otherwise hardly used. The reason for this is the faddish nature of the nonsense suffix *-sen*, discussed above as a component of *peazen*. 
Figure 5.3. Comparison of relative monthly frequency of two variants of peace with orthographic <z>.

An examination of -sen forms in the corpus confirms that wordforms featuring this suffix did not last. Moinsen, for example, declines sharply in usage from an early peak in 2000 to its last usage in December 2007. Bis densen is last used in mid-2007, around the time of the last usage of peazen. Auf jedensen (auf jeden 'of course' + -sen) is discarded by mid-2004, oh mansen (oh man + -sen) in mid-2006. Josen and yosen, unfortunately, see no further use after the conversation thread excerpted in (2a-b) above, demonstrating that not every neologism catches on, especially if it plays on a construction (x + -sen) which is already losing favor.

As for the more gradual decline in the use of the form peace, it might at first be tempting to compare this decline directly with the decline in battlen/batteln, but because the latter is a content word referring to a specific activity, rather than a discourse marker like peace, such a comparison is problematic. On the other hand, comparing the decline in the use of peace as a discourse marker, in this case a common leave-taking, with the decline in the activity of the hip
hop battle makes more sense; both were at one point practices indexing a tie to hip hop culture, but in both cases, the practices have become less prominent within the community. A concomitant possible factor is that in forums, as in other genres of CMC, norms of message composition do change over time; it could be that increased familiarity with the forum format over time has caused leave-takings more generally to fall out of favor, but testing this hypothesis would be complex and falls outside of the scope of the current analysis.

In summary, the variation found in the formulaic leave-taking *peace* appears to be a stable situation with heavy preference given to the English orthographic form *peace*. The form of the variants *peaz* and *peaze*, which are relatively frequent when compared against other alternatives, are best explained by the adoption of the English value */z/* for orthographic <z>, but with a twist: the */z/* follows the German phonological rule of word-final devoicing, yielding the form [pi:s] in both cases. Essentially, the application of a combined set of English-German orthographic-phonetic norms and phonological rules allows for a more diverse array of alternate, stylized forms—<peaze> appears nowhere in an 18-million word corpus collected from an American hip hop forum. In the next section, we shall see the most extreme example of orthographic variation in terms of broad application, but as in the previous cases, the explanation behind the orthographic forms is more complex than a simple reproduction of English orthography.

5.6. Syncretic spread: the participial (and non-participial) –ed suffix

The primary argument set forth in this section is that while an orthographic <-ed> form is established as a lower-frequency alternative to the German orthographic <-t> suffix in participles constructed from borrowed English verbs (examples from a Google search include: *gekicked, gephotoshopped, gejailbreaked*), there is, in the German hip hop community, a major extension of this <-ed> orthography to syncretic non-participial morphemes realized as -t in German. In particular, <-ed> is extended to borrowed English verbs in the 3rd person singular present tense and the 2nd person plural imperative. While I suggest that the extension of this orthography to the syncretic –t suffix used for 3rd person singular and 2nd person plural forms is largely peculiar to the hip hop community studied here, the primary goal of this section is to establish the existence of this orthographic form and its environments of use. The investigation of forms ending in –ed
was spurred by *released* and *produced*, two wordforms appearing along with the most frequent anglicisms in the corpus, listed in Appendix A.

5.6.1 Participial use of *-ed*

The borrowed verbs *released* and *produced*, and a large number of others like them, present an interesting orthographic innovation when surrounding context is considered. Examples (3a-b) demonstrate the use of these verbs as past participles in the perfect tense with the auxiliary verb *haben*, while (3c) demonstrates the use of the participial form in a passive construction:

(3a) [Post#020569, April 2006]

[...]
Genau das ist, Azad war früher bekannt dafür das an seinen Alben fast alles selbstgemacht ist, auf dem neuen Album glaub ich ein Lied das er **coproduced** hat, toll was Mehrverkauf mit sich ziehen muß.

'That's exactly it, Azad was known earlier for the fact that he did almost everything himself on his albums, on the new album [on the other hand], I think there's a single song that he **coproduced**, crazy what mass sales bring with them.'

(3b) [Post #000286, November 2005]

[...]
Zum Thema: Gänsehautfeeling war als Samy aufm Splash 04 Sneak Preview **performed** hat, wie die Menge abgegangen is

'On the topic: the goosebumps feeling was when Samy [Deluxe] **performed** Sneak Preview [a song] at Splash [hip hop festival] 04. how the crowd went nuts.'

(3c) [Post #000848, November 2005]

[...]
Die jetzigen Sachen die dort **released** werden (Candy Shop is das beste Beispiel) werden nur produziert um kommerziell erfolgreich zu sein.

'The current things that are being **released** there (Candy Shop is the best example) are only produced in order to be commercially successful.'

In (3a-c), German morphology would regularly form these past participles by taking the prefix *ge-* and the suffix *-t*. For example, the German infinitive *tanzen*, 'to dance' becomes the past participle *ge-tanz-t* 'danced'. In the perfect tense, this goes with the auxiliary verb *haben*, in this case *hat*. The use of the nativized fully German-inflected form is standard for some borrowings;

37 Here, as elsewhere in the dissertation, material excerpted from forums and in citations which I have added, generally in order to provide additional information, is indicated with square brackets [].
for instance, gerappt (the past participial form of 'rapped') is found 1,028 times in the corpus, rapped 45 times, and gerapped only 5 times.

By contrast, *geperformt / *geperformed, *gereleast / *gereleased, and *geproduct / *geproduct / *geproduced are not found at all, while the forms performed, released, and produced are readily found in the MZEE corpus. A probable explanation for this pattern is that these verbs have accent on a non-initial syllable, like German probieren, 'to try', bearbeiten 'to deal with something', or buchstabieren, 'to spell'. These verbs form their past participles without the addition of the prefix ge-, i.e., probiert rather than *geprobiert; bearbeitet rather than *gebearbeitet; buchstabiert rather than *gebuchstabiert.38

In the case of the -ed suffix, the explanation is a bit more complex. Some German irregular verbs keep their infinitive suffixes (-en) when they become participles, but -ed is not an infinitive suffix in English, nor a German suffix at all. -ed, however, is the English past tense suffix pronounced variably as [d]/[t] in English, and, crucially, would be pronounced by most German-dominant speakers as an unvoiced [t], the same pronunciation as the German past participial suffix -t. This outcome is the result of a productive phonological rule known as Auslautverhärtung (‘final sound hardening’), or as it is known to English speakers, word-final devoicing, discussed in previous sections in its application to the phoneme /z/. Word-final devoicing affects both fricatives like /z/ and stops like /d/. This circumstance means that -ed, or rather, <-ed> is a felicitous borrowing orthographically: a German-speaking forum user may type <released> or <releast>, without the spoken forms being especially different for most German speakers.

An important note at this juncture is that this use of -ed with borrowed verbs in participial form, while perhaps more common in the hip hop community, is by no means unique to this subculture. Newer borrowings in more mainstream usage also exhibit this variation—for instance, a Google.de exact word search for the following forms shows that –ed is a less frequent but existing option for participial formation of the verb google:39

38 My thanks go to Hans Hock for assistance with the stress-based analysis.
39 Reported searches were performed on google.de 12 November 2011.
<table>
<thead>
<tr>
<th>Wordform</th>
<th>Google.de frequency</th>
<th>Wordform</th>
<th>Google.de frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>goooglet</td>
<td>~4,650,000 hits</td>
<td>goooglet</td>
<td>~949,000 hits</td>
</tr>
<tr>
<td>googeld</td>
<td>~19,000 hits</td>
<td>googled</td>
<td>~152,000 hits</td>
</tr>
</tbody>
</table>

Table 5.6. Comparison of participial forms of the borrowed verb google.

It is also worth noting that google as a verb falls into the class of borrowed verbs with orthographic variation induced by a stem-final <-le> or <-re>, like batteln/battlen; however, unlike batteln/battlen the orthography which keeps the English stem google intact (in the third column of Table 5.6), appears to be the disfavored one in this case, except for forms ending in d. I suggest here that this is a result of the orthographic form <googeld> retaining neither the English stem google faithfully, nor the -ed suffix; alternatively, this could be a result of the fact that the form google can be considered 'native' to written language; percentage-wise, one would expect more of the instances of interaction with google to be written and read, rather than spoken and heard, like battle. Finally, (3c) above in particular is worth a discussing, as it contains both released and produziert, 'produced', for which we know from (3a) the alternative produced is possible. My explanation for this is twofold—first, the use of these forms seems to be in variation, even within speakers. I would not expect most speakers to use one or the other form exclusively. Second, reading from context, this use of produziert may refer to a more generic production of something by a record label, in this case, hip hop songs, rather than the specific notion of produced which refers almost exclusively to the discipline (accomplished by a Producer) of putting together beats, melodies, and samples for rappers.

5.6.2 Non-participial use of –ed

The most exceptional thing about the -ed form in the MZEE corpus is that in the context of the German hip hop community (and to my and my native informants' knowledge, nowhere else), this orthographic -ed is not limited to participial forms:

(4a) [Post #004014, December 2005]
[...] ich glaub die streets albums haben hier auch die meisten im schrank und dass dizzee sich selbst produced ist auch nicht so neu.

'I think most people here also have the streets [a hip hop group] album in the closet and the fact that dizzee produces / ?produced himself [his own songs] is also not so new.'
Sido is now so successful because he raps / *rapped well. Clearly at the beginning his rap was still somewhat "rough", but since then he has developed himself better. Already a crass [colloquial usage, positive meaning] guy.

PeaZe

This thread sucks / ?sucked

Oh wow: Moses raps in English now and again!
Or should I have.. written ['']speaks rhythmically now and again[''].
this is a bullshit thread!
Just do what you want, a lot is influenced by the U.S. of A anyway.'

In (4a-4c), interpreting the wordforms (produced, rapped, sucked) as past tense is grammatically possible, assuming a full codeswitch to English morphosyntax—this is required because no auxiliary verb is present, a prerequisite for the use of a German participle. The interpretation of these forms as past tense, however, is not likely in the contexts given here. The only other plausible reading (and the one favored by the context in these three examples) is as a simple present tense reading, which in all three cases would require the 3rd person suffix -t (a common and syncretic suffix in German, as will be shown in (5). In fact, in the case of (4c), a debate begins in the discussion thread between the author of that post and a detractor of his or her usage who suggests the orthography <sukt> on account of the obvious impossibility of borrowing grammar from English—the author of (4c) defends his or her choice, noting in a later post that <sukt> sieht total komisch aus, 'looks totally weird' and even later, that <sukt> sieht doof aus, 'looks dumb'. The excerpted discussion is reproduced in German and in translation in Appendix B.
As mentioned above, a past participial reading is impossible in all three cases because of the lack of auxiliary verb \((\text{haben/sein})\) in each case. The other German past tense, the preterite, is also out; German verbs in the preterite gain a \(-t\) suffix that appears before the standard suffix for person/number agreement, or else undergo ablaut, a process whereby a word-medial vowel changes in quality, e.g. Ger. \(\text{fliegen} \) 'to fly', preterite \(\text{flog}\)—although note that ablaut would not be expected in the verbs in (4a-c). This process is shared by certain English verbs, e.g., \(\text{freeze-froze}\), but neither ablaut nor the requisite person-number markers after the \(-t\) suffix are present in (4a-c). The next example shows a definite extension of the orthographic \(-ed\) suffix to another German form marked by a syncretic \(-t\) (or \(-et\)): the second person imperative plural.

\[(5) \quad \text{[Post #012496, February 2006]}
\]

\[\ldots\] Also alle deutschen Producer vereinigt euch, brüllt eure Namen wie Premo am Anfang des Beats, \textbf{sampled} Trademarks, gebt alles.

'So all you German producers come together, yell your names like Premo [DJ Premier] at the beginning of the beats, \textbf{sample} / *\textbf{sampled} trademarks, give it your all.'

The occurrence of \textit{sampled} as one of several coordinated second-person plural imperative verbs \((\text{vereinigt, brüllt, and geht})\) makes it clear that the English orthography \(<-\text{ed}>\), which is limited to the past tense in written English, has taken root in this German-language hip hop forum as an orthographic representation of the German suffix \(-t\) which is syncretic in German, realizing one of several morphological categories (past participle, third person present singular, and second person imperative plural). The next question, then, is whether this phenomenon is specific to this forum. External evidence from YouTube suggests that this is not the case, as seen in Figure 5.4:
Figure 5.4 is a screen capture from the German-Turkish collaborative hip hop video 'Alles Tamam' (Ger. *alles*, 'everything' + Turk. *tamam*, 'OK'). The video's uploader exhorts the viewer, both in the red pop-up bubble and the video description, to vote for the clip on MTV Urban:

(6a)  [pop-up bubble text, German]

\textit{voted bei MTV Urban für diesen Clip>>>>>link rechts!!!}\!

[URL]

\textit{'vote [3.pl.imp] for this clip at MTV Urban>>>>>link to the right!!}'


The first and last examples here contain essentially the same second-person plural imperative, and, crucially, the repetition of this message with minor rephrasing of the text demonstrates that this is not a one-off typo which has been cut-and-pasted. The Turkish translation is essentially a trilingual hybrid form: vote is an English root borrowed into German, given German infinitive morphology (-en), and then paired in this Turkish sentence with yapın, the Turkish polite imperative 'do (something)', i.e. voten yapın means something like 'please vote!'.

Thus far we have evidence for the nativization of the orthographic form –ed, and crucially, its extension to use as a grammatical 'false friend' for third person singular and second person imperative forms, from a hip hop discussion forum and an unknown YouTube video uploader's comments. Does this orthographic form have more widespread currency, and is there more evidence that this is not an elaborate and unlikely series of typos? Another YouTube video provides further evidence. Roey Marquis II, a German hip hop producer/DJ features two of his songs, Pickel, with Olli Banjo rapping, and Tschukka with a series of rappers in a double-video on his official YouTube channel. A portion of Olli Banjo's verse follows:

(7) Greif' das Mic, doch verbrenn' dir deine Finger net am Internet, denn wenn man Behindertrap aus dem Zwinger lässt, Da flowt nix, wenn ein Kleinkind das Mic nimmt, das Internet einnimmt und die Scheisse dann eintippt.

'Grab the mic, but don't burn your finger on the Internet, because when one lets handicapped-rap out of the dungeon,
It doesn't flow [lit. there flows nothing], when a toddler takes the mic, engages the Internet, and types the shit in.'

The lyrics, as available from third-party lyrics host magistrix.de, transcribe the anglicism flow with the phonologically and orthographically native German -t 3rd-person singular present suffix, and while phonetic analysis is inconclusive on account of the interference of the beat (see Figure
5.5), it seems likely, on account of the preceding vowel quality, that the segment is an unreleased [t].

Figure 5.5. Spectrographic representation of the sequence flowt ni[x] from the song Pickel.

The case here would be closed in favor of straightforward assimilation of the English verb *flow* as the German form *flowt* in an orthographic sense, but for the fact that the video (which uses paper masks as a motif) features several masks with the lyrics printed on them. In a rapid video cut, 'DA FLOWED NIX' clearly appears across the screen, as in Figure 5.6:

---

42 Thanks and credit go to Vandana Puri Sharma for assistance with the spectrographic representation of this example in *Praat.*
This visual use of what I call the ‘non-participial –ed’ is significant because it demonstrates the reach of this orthographic form as not something used only by hip hop fans but also by a well-established hip hop artist, Olli Banjo. The case of the non-participial –ed thus provides further evidence that the borrowing of English verbs and their attendant participial morphology (a case where the difference between English and German participial morphology is neutralized by the operation of word-final devoicing) can lead to wider orthographic variation. In the following

---

43 'Roey Marquis II feat. Curse, Germany & Italo Reno - Tschukka'  

44 While the title of this YouTube video doesn’t indicate it, it is in fact a double-video for two songs off of the album ‘Herzessenz’ by producer Roey Marquis II; The first of the two videos is for the song ‘Pickel’ featuring Olli Banjo.
subsection, I turn briefly to a quantitative analysis of the –ed suffix and its use in participial and non-participial contexts in the MZEE corpus.

5.6.3 The distribution of participial and non-participial –ed
As with the plural -z and the forms of peace in sections 5.4 and 5.5, I present here graphs of the usage of participial forms on a monthly basis throughout the MZEE corpus. As this analysis involved the discovery of any English-derived verb ending in –ed, the situation was more complex than a simple word search. First, all words in the corpus ending with –ed were located, and a list of stopwords containing names like khaled, and adjectives like stoned or wicked were ignored. Likewise, all words ending in -ied and -eed were ignored, which removed a number of common German words like Lied 'song' or Unterschied, 'difference' and common English nouns like weed and seed. Next, the remaining words which were adjacent to an auxiliary verb, i.e., were located with an auxiliary verb (like haben, hat, hast, hatte, sind, ist, war, werden and so forth) either immediately preceding or within the next two words, were divided into the classes 'participle with ge' (if the word begins with ge-) and 'bare participle' (if, like produced and released, no prefix ge- was present). The remainder were placed into another file (roughly 4,000 words) and annotated by hand as third-person singular present, second-person plural imperative, or false positive/indeterminate. In order to correctly and accurately annotate these forms, four words of surrounding context (two on each side) were presented in a fashion similar to a concordance. The results of these analyses revealed that the corpus contains 2,163 instances of –ed verb forms used as participles, of which 1,436 are prefixed with ge- and 727 are without prefix; additionally, the corpus contains 886 non-participial uses of -ed, of which 746 are third-person singular present, and 140 are second-person imperative plural. In a 12+ million word corpus, these figures are not especially large, but the -ed forms are used consistently throughout the corpus, as is seen for the participial monthly frequencies (Figure 5.7) and the non-participial monthly frequencies (Figure 5.8):
Figure 5.7. Relative monthly frequency of -ed forms prefixed with ge- (light gray, dashed line), and bare -ed forms (dark gray, dotted line).

Figure 5.8. Relative monthly frequencies of third-person singular present non-participial usages of -ed (light gray, dashed line) and second-person plural imperative usages of -ed (dark-gray, dotted line).
In Figure 5.7, it is noteworthy that participial -\textit{ed} forms seem to slowly increase in usage over the timespan of the corpus; this indicates that these forms have become more accepted over time in the community. Because of the small number of examples used in creating Figure 5.8, it is inadvisable to read too much into the general slope of the line; suffice it to say that both 3\textsuperscript{rd} person singular present forms and 2\textsuperscript{nd} person plural imperative forms are found throughout the corpus.\footnote{The corpus is sparse in terms of data quantity for several months in 2001. Combined with the general burstiness of the data, this accounts for the dip in many of the charts in this chapter during that year, a dip that is most evident for low-frequency forms.} The difference in overall frequency between the two types of -\textit{ed} forms in Figure 5.8 is also not surprising, and is likely the result of 3\textsuperscript{rd} person singular present being a more common form overall.

5.7 Conclusions

In this chapter, I have presented an analysis of the linguistic features and diachronic distribution of several wordforms selected from a list of frequent anglicisms in the MZEE corpus. These wordforms were not solely selected based on frequency, but rather because of unexpected linguistic properties that the words, and their variants, displayed in the corpus. The results of these analyses, in combination, primarily support the argument that the nativization and adaptation of borrowed words are not always straightforward, i.e., that a binary distinction between unassimilated loanwords and fully Germanized loanwords is essentially untenable, and that both English phonological factors (like [z] for <z>) and morphological factors (the orthographic form, if not the meaning, of the morpheme -\textit{ed}) and German phonological factors (like the word-final devoicing rule) and morphological factors (the syncretic suffix -\textit{t}) make a significant contribution to the word's orthographic form. Also at work in these examples is a phenomenon known as \textit{hyper-foreignism}: Hock and Joseph (1996: 270) cite cases like the substitution of [ž] for [j] in the South Asian \textit{raja(h)}, 'ruler/monarch/king', upon its borrowing into English and the dropping of the final [s] in the English pronunciation of French \textit{coup de grace}, 'mercy-killing blow', noting that these cases of alternate pronunciation do not follow the rules of, e.g., French in the latter case, but rather English speakers' perceptions of those rules. Another case cited by Hock & Joseph, this time of hyper-foreign morphology rather than pronunciation, is that of the morphological addition by English speakers of (generally) -\textit{o} to Spanish forms like \textit{problemo}; Hock & Joseph (1996: 270) comment:
At work in these cases is a phenomenon we have seen before and will see again – most speakers are not linguists. What matters are ordinary speakers’ perceptions of what makes a word or sound seem foreign, not what the actual facts of the foreign language are. These facts are for linguists to worry about; speakers are too busy using their own language to be concerned with such fine details.

The cases discussed in this chapter, and especially that of non-participial -ed, exhibit some parallels to this observation. While the key issue with <-ed> is that it is an orthographic representation of a German morpheme, rather than the English past-tense morpheme, its orthographic form is certainly conditioned by German speakers' perceptions of what looks English, rather than by a notion of adherence to English grammatical rules.

It is also a primary goal of this chapter to establish the importance of orthography in revealing the linguistic factors at work in the borrowing process—without varied orthographic forms, for example, knowledge of the role of word-final devoicing in borrowing would not be accessible, nor would it be evident that <peazen> ~ /pi:sn/ is composed of peace + -sen rather than peace + -en. Finally, it is important to note the role that culture plays: while the combination of English and German linguistic rules allows for these varied forms, it is, as Androutsopoulos (2000) argues, the stylistic prerogative of individual speakers which ultimately motivates their use and, to use an example from this analysis, leads to the adoption of new forms like non-participial -ed.

While the participial usage of -ed is found in German contexts outside of hip hop culture, the non-participial usage of -ed with borrowed verbs in German is thus far unattested outside of the online hip hop community, even in other domains, like advertising, rich in English borrowings. Until contrary evidence is presented, I claim here that this orthographic phenomenon is unique to the hip hop community and may be related to the loosened orthographic norms of many online modes of communication. However, it is important to be circumspect in claims of novelty and the attribution of a special or mysterious linguistic power to new media, as Stein (2005: 201) notes:
[...] the well known claims in cultural theory about the hybridity of the modern media, particularly the internet, show little knowledge of the nature of spoken and written communication. What they do is fall prey to the same unrealistic segregationist views of much modern theorizing about language [...] In other words, if you do not see the hybridity of normal communication in terms of the sources of meaning in context and how they are constructed ad hoc, then, of course, the new media would appear to be particularly hybrid. Thus a myopic view of the polyphony of information sources causes modern constructionists to overlook the constructional character of spoken and written communication, and construct a non-existing contrast to the new media.

As Stein suggests, it is not the difference between written and spoken forms, i.e., orthography and morphophonology, that is crucial here; rather, it is important to recognize that spoken and written communication have not developed separately since the introduction of orthographic technology, but rather that norms and rules from each side have always influenced one another; an existing process of transfer between the two realms is merely accelerated in an age of computer-mediated discourse.

To conclude, it seems that the US-oriented hip hop community in Germany, a product of globalizing forces, functions as a conduit for a wide range of English borrowings, but that rather than straightforwardly reproducing American forms, these forms, like the orthographic -ed, are taking on a life of their own in German, not strictly following the regulations of their origins but rather fusing with local meanings, styles, and orthographic norms. As for major social concerns about the future of the German language, these cases demonstrate that borrowed orthographic variations (like -ed for -t in non-participial forms) are conditioned not only by English morphophonology, but crucially by German morphophonology as well. These seemingly alien forms, like the orthographic variation in verbs ending with –ed and the <z> in Beatz and Peaz are in fact made more likely due to the German rule of word-final devoicing. While concern for the future of the German language is present in the media and the political sphere, most linguists would take the large speaker population as an indication of the German language’s continued well-being—and research like this suggests that borrowing into German is in fact proceeding not haphazardly but in a rule-governed and systematic fashion, constrained by German linguistic rules. Finally, this research illuminates the way in which linguistic material is borrowed from English to German, and crucially, how borrowing interfaces with orthography, and can at times exploit orthographic norms. Orthography is often overlooked in studies of linguistic borrowing,
but it has valuable things to tell us about the way borrowing works—and can illustrate changes that are not evident in spoken language, but are nonetheless conditioned by the rules of spoken language.

Having examined specific anglicisms and their morphological-orthographic nativization in this chapter, I have only begun to leverage the opportunities provided by a large, decade-long natural language corpus. I turn next to a computational analysis of the ways in which the fate (as determined by relative frequency change) of larger sets of anglicisms over time might be predicted by various factors, an analysis that will also reveal more about how the behavior of borrowed forms in the MZEE corpus differs (sometimes greatly) from the behavior of native German forms.
CHAPTER 6
PREDICTING CHANGES IN ANGLICISM USE

Ey hör mal gut zu, auch ohne Product Placement und Marketing sind wir auf Straßen King weil Kids unsere Phrasen singen. Hip hop ist Massending, megatrendy wie 'n Nasenring, Hör Tassen klingen Backstage weil vor der Bühne Massen springen.

Hey listen up, even without product placement and marketing we're king on the streets 'cause kids sing our catchphrases. Yeah hip hop's a thing of the masses, mega-trendy like a nose ring, hear the sound of toasting backstage because in front, the crowd's jumping.


6.1 Introduction and motivation

German hip hop fans' choice to use English borrowings can be considered, in essence, a problem of contact-induced linguistic variation leading to contact-induced linguistic change. There is, after all, nothing inherent in the German language which would prevent the establishment of new native wordforms or calques for hip hop-related concepts which are generally expressed through anglicisms like dope, wack, diss, etc. In fact, a number of native German words have functioned as hip hop-related terms or, more commonly, extended semantically to encompass hip hop-related meanings, e.g., derb '[coll.] rough [positive]', krass '[coll.] crass [positive]', and alter/digga '[regional] dude'. Establishing this as a problem of linguistic variation might suggest that the problem lends itself to variationist explanations, whereby the variation would be examined in terms of its relation to classic sociolinguistic variables, such as age, ethnicity, gender, and so forth. However, the dataset examined here (i.e., the MZEE corpus) does not particularly lend itself to such analysis because it is demographically lean, as explained later in this chapter. In addition, while some anglicisms in the corpus may be in variation with one or more German wordforms, this is not the case for many of the loanwords, which have either won out over nascent German alternatives, or for which German alternatives never existed.

Whereas the classical variationist paradigm may not be particularly well suited to answering questions about anglicisms' success and failure in this case, there are other statistical methods, used in the present chapter, that can illuminate factors that do play a role in shaping the language.

of the German hip hop community—which is to say that these factors condition the frequency (and change in frequency) of anglicisms diachronically in the corpus. I begin this chapter by examining the demographics of the MZEE corpus insofar as such information is available, in order to at least establish a general picture of the hip hop community as instantiated in the MZEE.com forums. For the remainder of the chapter, I establish the fate of anglicisms in the German hip hop community and to examine several factors hypothesized to contribute to, and in turn to serve as predictors of, the success or failure of English borrowings in the German hip hop community, thereby gaining insight into the dissemination of borrowed forms through shared-interest-bound communities of practice.

Having developed and described an anglicism classification system for the MZEE corpus in Chapter 4, I turn in the present chapter to analyses made possible by the automatic identification of anglicisms in the text. Access to the 850 English-derived wordforms which occur over 100 times in the corpus (reproduced in Appendix A) allows for a more powerful, computational investigation of the use of these wordforms throughout the corpus. In this chapter, after discussing the demographics of the MZEE.com forums in 6.2, I will present the relative frequency of identified anglicisms in the MZEE corpus over time in section 6.3. Next, I turn to the research question of which factors play the greatest role in a loanword's ultimate success, maintenance, or extinction, first comparing the effect of the incidence of wordforms in the English-language (Covo) corpus on frequency change in the MZEE corpus in section 6.4. Second, I analyze the relationship of word frequency and dissemination in MZEE to frequency change for anglicisms and other words. This last analysis is in large part inspired by the work of Altmann, Motter, and Pierrehumbert (2011) and a concomitant presentation of that research by Pierrehumbert (2011), the relevant portions of which will be discussed in section 6.5. Altmann et al. performed a number of diachronic analyses on two corpora drawn from Usenet discussions in the groups comp.os.linux.misc and rec.music.hip-hop, which respectively consist of online discussions revolving around the Linux operating system and hip hop music. They found that word niche, operationalized as two measures of dissemination (dissemination over users \( D^U \) and dissemination over threads \( D^T \)) is more strongly predictive of a word's long-term fate in these corpora than a control measure, frequency \( f \). These findings are intriguing because they suggest that complex contextual factors are ultimately more important to a word's success or failure than
a word's initial frequency. If these factors play a large role in the uptake and dissemination of novel words in monolingual English discussions, it could then be hypothesized that dissemination over users and threads would likewise play a larger role than frequency in the fate of anglicisms—but the results of a similar analysis on the MZEE corpus challenge that hypothesis, as frequency change for anglicisms is found to be more strongly predicted by initial frequency than by either dissemination measure. In the concluding section of this chapter, I contrast these results with those of Altmann et al., suggesting that the primary factor which leads to the success of anglicisms in the MZEE corpus is in fact low initial frequency, and conversely, that anglicisms' failure can be attributed, at least in part, to high initial frequency (a result corroborated in part by the findings of Chesley & Baayen 2010).

6.2 The demographics of the MZEE corpus

To investigate sociohistorical questions in any collection of linguistic data, both the inherent qualities of the speakers (in terms of classic sociolinguistic variables as mentioned in the introduction) and community norms must be taken into account. In Iorio's (2009: 129) study, he designates the sociolinguistic community of the massively-multiplayer online game 'City of Heroes' a demographically lean community, a term referring to “online communities where the association between a member's offline demography and their online representation [is] either incongruous or absent.” While forum profiles on MZEE.com provide a space in which to share voluntary demographic information, this information, when provided, is often incomplete or of varying integrity. The sociolinguistic variable of age is a case in point: while some users list their full date of birth, allowing their age to be calculated for any given forum post, some list only the day and month (but not the year) of their birthday, or list only their age, which is not automatically updated by the system with the passage of time—and as such would represent their age at the time of filling out their profile—adding an undesirable amount of variability to this information.

To glean what demographic information might be available about the users of the MZEE.com forum, a sample of 50 user profiles was collected from the most recent pages of the MZEE.com forum thread titled Was macht jetzt..., 'what are they doing now?', a general-interest thread where interested users can ask the community about the recent activities of previously popular artists.
The thread was selected because it attracted a large number of participants, and yielded a diverse, if not entirely random, sample of users both new and old.

Of the 50 user profiles investigated, 16 provided their purported age, and of these, 10 provided a year of birth (in which case the system automatically calculates the user's age at the time profile information is accessed; in this case, March of 2011. The reported ages ranged from 21-31, with a mean user age of 24.9. Of the 26 out of 50 users who reported gender information, all (100%) reported 'male'. The sampled users registered their profiles between 2001 and 2010, with a mean date of registration in September of 2007. Users' post counts were also provided by the system; the users sampled ranged (in March 2011) from 1 post to 23,865 posts. Three of the users sampled were moderators in the community; these accounted for the three highest post counts—i.e., the moderators tend to be the largest contributors on the forum. The mean post count among the users sampled was 2,962 posts.

Short descriptions of the users' avatars—small pictures appearing under users' names—were also collected: 38 of the 50 users had uploaded such an avatar. Several trends became clear: six of the avatars were the covers of hip-hop albums, four were related to soccer (football) teams or players, three were suggestive pictures of women, and, perhaps in a nod to the anonymity of the medium or the criminality often associated with hip hop, three avatars involved figures wearing masks of some sort, gasmasks and kerchiefs covering the face being the most common.

To summarize, those users who chose to share some sort of demographic information overwhelmingly reported themselves as males in their 20s. The self-reported values may, of course, be spurious: as with many forms of asynchronous computer-mediated communication, there is no accountability system in place on the MZEE.com forums to ensure that someone is who they say they are. In fact, there may be significant disincentive for female users, or users significantly below or above the average age, to include such demographic information. In the case of gender, Herring (2003: 206) notes that in many online modes, women “can present themselves so as to minimize discrimination and harassment” by refusing to self-identify as female. It is not unlikely that in the overtly macho, competitive, and sometimes misogynistic culture of hip hop as instantiated on the MZEE.com forums, female fans routinely neglect to
provide information or cues about their gender. Likewise, fans younger than 20 or older than 30 might routinely neglect to provide information about their age; younger fans might be self-conscious about being seen as immature 'kids' and find it harder to win respect, and older fans might be made to feel out of place or out of touch with the cutting edge of the culture. For these reasons, even the limited information available about this demographically lean community should be taken with a grain of salt.

To provide a more complete picture, I would suggest several valid inferences which can be made with regard to the users of the MZEE.com forums without reliance on self-reported data. First, it is fairly safe to say that the vast majority of those posting on the MZEE.com forums are German speakers, and that they share an interest in hip hop, as there are few, if any, logical incentives for users who do not fit this profile to participate. The predominance of the German language on the forums and the subject of discussion would quickly yield diminishing returns for non-German speakers and those uninterested in hip hop music or culture.

Despite the likelihood that these inferences are correct, they do not allow the researcher to easily divide the users of the MZEE.com forums into neat sociolinguistic categories, and the unreliability and incompleteness of demographic data, as previously discussed, precludes traditional variationist analysis. However, the sheer plenitude of the data, combined with the length in real time of the corpus (over a decade) give the researcher greater options in terms of computationally tractable diachronic analysis.

6.3 Prevalence and overall patterns of anglicism use in the MZEE corpus

In this section, I discuss several basic diachronic analyses, including the size of the MZEE corpus over time, and the diachronic frequency of those anglicisms identified in Chapter 4.

6.3.1 The MZEE corpus over time

To provide a baseline for over-time analyses, it is necessary to consider the total number of words or users in each window of time under consideration; if this is not done, and the absolute frequency (not relativized to the corpus size within a time window) is taken instead, a word
occurring at chance could seem to trend up and down as the total number of words increases and decreases, when its relative frequency at any given time $t$ is actually fairly constant. With this in mind, monthly and yearly word totals were prepared over the span of the corpus. The entire corpus is roughly 12.5 million words in length (after the pre-processing described in Chapter 3), comprising about 382,000 unique posts spanning the time period from March 2000 to March 2011. Figures 6.1 and 6.2 show a representation of the corpus (in words) on a monthly and a yearly basis, respectively, over this time span.\footnote{The number of posts per day/month was also analyzed; trends were nearly identical, indicating that the average number of words per post remained fairly constant.}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6.1}
\caption{MZEE corpus word totals per month.}
\end{figure}

From these data, it is apparent that the forum was most active between late 2002 and early 2007. There are several possible explanations for this: one, that the forum started fairly small, gaining steadily in popularity until about 2005, when the rise of social-networking sites like Myspace, Facebook, and StudiVZ (a German social-networking site) precipitated a drop in forum use.
Another possible explanation involves the meteoric rise of the controversial Aggro Berlin record label, which brought a harder sound to German hip hop culture and, through influential (and constantly feuding) artists like Sido, Bushido, and Fler, scored a large number of chart-friendly hip hop hits, popularizing the genre to a large number of young new listeners in the early-to-mid 2000s. Aggro Berlin, after several troubled years, folded in 2009. It is probable that both of these factors played a role in the relative popularity of the MZEE forum over time.

6.3.2 Prevalence of anglicisms in the MZEE corpus

Figure 6.3 shows the relative frequencies of all identified anglicisms (absolute frequency > 100; \( n = 850 \)) in the corpus on a monthly and yearly basis, revealing that the use of these frequent anglicisms seems to increase (relative to the use of other words) over time in the forum. However, it is crucial to note that 2011 data represents only the first three months of the year, and the amount of data considered varies considerably with the number of words each month/year (see above). Therefore, I refrain from making definitive pronouncements about the relative use of anglicisms over time, but even discarding the partial years 2000 and 2011, a general trend upward is visible. In Figures 6.3 and 6.4, below, values on the y-axis are indicative of the percentage of word tokens in the corpus per year/month that these 850 anglicisms account for.
Several tendencies can be gleaned from Figures 6.2 and 6.3 in terms of the use of these frequent anglicisms. First, that these anglicisms account for, in general, around 3-4% (average monthly percentage: 3.70%) of the word tokens in the corpus (recall that in section 4.6, the likely overall anglicism frequency in the corpus was estimated at 6-8%). The distribution of anglicism types,
judging by the candidate list, has a very long tail of low-frequency anglicism wordforms—59.65% of the candidates being *hapax legomena*, 13.42% occurring twice, and 5.89% occurring three times. In Figures 6.2 and 6.3, it seems that the use of the most-frequent anglicisms increases throughout the timespan of the corpus. The analyses presented in the next two sections seek to address the word dynamics which act on anglicisms (and other words) in the MZEE corpus.

6.4 Relating anglicisms between corpora

The use of the word *anglicism* rather than the more general *borrowing* presupposes a source language for the language material considered here, namely English. Given the nature of the domain under consideration, and taking hip hop as a globalizing culture centered in the US, it is expected that the use of anglicisms in this domain is in part conditioned by their use in US hip hop contexts, given that American hip hop is very popular in the German-speaking sphere, and that American musical releases and other hip hop-related media serve as primary texts for both American and German hip hop fans. In this analysis, I investigate the relationship between the use of anglicisms' source words in the US-based Project Covo forums and their fates in the MZEE corpus.

6.4.1 The aligned corpora

In order to investigate a diachronic relationship between words in two corpora, it is first necessary to ensure that the timeframes involved make sense, i.e., that I am not, for instance, trying to predict frequency change from 2000-2004 in MZEE from word frequency in 2005 in the Covo corpus given that influence is hypothesized to proceed forward in time. I examined the two corpora in terms of their annual corpus size, and stipulated a lower limit of 1 million words for each time window under consideration. In the end, a sequence of four consecutive year-long time windows was found in which both corpora provided at least 1 million words. These time windows were Nov. 2003 – Oct. 2004, Nov. 2004 – Oct. 2005, Nov. 2005 – Oct. 2006, and Nov. 2006 – Oct. 2007. For the sake of convenience, these will be referred to as the 2004, 2005, 2006, and 2007 windows, respectively. Final word counts for each year-long subcorpus are recorded in Table 6.1.
<table>
<thead>
<tr>
<th></th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covo</td>
<td>2,885,216</td>
<td>4,780,656</td>
<td>1,480,316</td>
<td>1,548,505</td>
</tr>
<tr>
<td>MZEE</td>
<td>1,833,669</td>
<td>2,488,620</td>
<td>1,737,561</td>
<td>1,010,906</td>
</tr>
</tbody>
</table>

*Table 6.1. Corpus sizes for annual subcorpora used in cross-corpus analysis.*

6.4.2 From wordforms to stems

As is evident from the variety of wordforms in Appendix A, and from the discussions of affixes and compounds in Chapter 4, anglicisms include uninflected English stems (e.g., *battle, rap, flow*) as well as English stems with (apparently) English inflectional or derivational morphology (e.g., *battled, rappers, flows*), English stems with German inflection (e.g., *gebattlet, 'battled'; rappern, 'rappers'; flowen, 'to flow*), and English stems with German derivational affixes (e.g., *battlemässig, 'battle-related'; rapperische, 'rapper-like'; flowendere, 'more flowing*), as well as compounds with one or more English parts (e.g., *battleraporientierter, 'someone oriented towards battle-rap'; hiphopgangstaghettorapper, 'hip hop-gangsta-ghetto-rapper'; maschinengewehrflow, 'machinegun flow*). In order to perform a proper frequency comparison of anglicisms in the MZEE corpus with their counterparts in the COVO corpus, it was necessary to account for these differences, i.e., for the greater diversity of wordforms in the German corpus due to the more highly-inflected nature and greater occurrence of word compounding in German. To this end, a set of 106 orthographic character strings corresponding to word stems (including, e.g., *<battl>, <rap> and <flow>* ) were identified from the 850 known anglicisms reproduced in Appendix A. The frequencies of anglicism stems in MZEE and Covo were determined for each time window by finding wordforms including these character sequences and summing their frequencies. In this process, it was also necessary to include a list of character sequences which disqualified a possible wordform from being counted towards a stem frequency, e.g. for *<rap>* , a number of spurious forms including the character sequence *<graph>* (*Biographie, Graphic, Photograph*, etc.) were excluded from the analysis. For each stem, frequency information was collected from the Covo and MZEE corpora for each year-long time window; log-frequencies were then calculated.
6.4.3 Results

In order to determine the fate of words in the MZEE corpus, I follow Altmann et al. (2011) in using the change in base-10 logarithmic relative frequency ($\Delta \log_{10} f$) from one time to another. Log transformation is used because, as noted by Gries (2010: 272), it “has the effect that the otherwise very skewed distribution of word frequencies is linearized and allows us to use simpler linear correlation measures to compare the corpus frequencies with other data.” While log-frequency at a later time alone (rather than change in log-frequency) has been used as a response variable (the value to be predicted) by, e.g., Chesley & Baayen (2010), change in log-frequency has the advantage of being relativized to a word's earlier frequency, and allows for comparison of the present results, particularly in the dissemination analysis presented later in this chapter, with those of Altmann et al. (2011).

In this analysis, the predictor variable under consideration—which predicts change in log-frequency in the MZEE corpus—is log-frequency in the Covo corpus at an initial time window $T_i$. To gauge the effect of Covo frequency on change in MZEE frequency through different timespans, MZEE frequency change was calculated for all pairs of years in the analysis (i.e., 2004-2005, 2004-2006, 2004-2007, 2005-2006, 2005-2007, and 2006-2007), and data for all comparisons using timespans of equal length were aggregated. This yielded three comparisons total, namely $\log_{10} f$ at $T_i$ with $\Delta \log_{10} f$ at $T_{[i: i+1]}$ (timespan length 1); $\log_{10} f$ at $T_i$ with $\Delta \log_{10} f$ at $T_{[i: i+2]}$ (timespan length 2); and $\log_{10} f$ at $T_i$ with $\Delta \log_{10} f$ at $T_{[i: i+3]}$ (timespan length 3).\textsuperscript{48}

Correlations were calculated for these comparisons and all three were found to be significant; these results are reported in Table 6.2, with scatterplots of these correlations presented in Figure 6.5. $R$, the correlation coefficient, reports the strength of the relationship between the sets of data. The $p$-value is also reported, which expresses the significance of the relationship. $R^2$, the coefficient of determination, is a measure of how well the predictor variable predicts the response variable, specifically the percentage of variance in the response which is explained by the predictor variable. The $N$ reported here refers to the number of data points considered; degrees of freedom $df$ is easily calculated as $N-2$.

\textsuperscript{48} Variation of the start distance of the response timespan from the initial time window, e.g., $\log_{10} f$ at $T_i$ with $\Delta \log_{10} f$ at $T_{[i+2: i+3]}$ was also performed, and results of those correlations were significant but weaker than those reported here.
Table 6.2. Correlations between lemma frequency in Covo and lemma frequency change in MZEE for indicated timespans.49

<table>
<thead>
<tr>
<th>Timespan length</th>
<th>$R$</th>
<th>$p$</th>
<th>$R^2$</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1891</td>
<td>0.0007</td>
<td>3.6%</td>
<td>318</td>
</tr>
<tr>
<td>2</td>
<td>0.3130</td>
<td>&lt; 0.0001</td>
<td>9.8%</td>
<td>212</td>
</tr>
<tr>
<td>3</td>
<td>0.2327</td>
<td>0.0164</td>
<td>5.4%</td>
<td>106</td>
</tr>
</tbody>
</table>

Figure 6.5. Scatterplots of Covo frequency vs. one, two, and three-year change in MZEE frequency.

It is worth mentioning at this point that language use is conditioned by numerous complex social and contextual factors, many of which are difficult if not impossible to capture quantitatively. Finding that a quantitative predictor variable, then, is significantly correlated with and explains

49 Here and in subsequent tables, significant correlations ($p < 0.05$) will be listed in italics.
at least some portion of word frequency change (even if the correlation is fairly weak) is an interesting result.

Returning to Table 6.2, if we consider the change in $R^2$ between the different timespan lengths, we see that frequency in the Covo corpus predicts frequency change in MZEE best in a two-year timespan. Because the time windows considered here are measured in real time, this result suggests a time lag of roughly 2 years in the effect of high (or low) frequency in Covo and a subsequent increase (or decrease, respectively) in frequency in the MZEE corpus. These results indicate that the frequency of an anglicism in German is in part determined by its (earlier) popularity in English.

6.5 Dissemination of anglicisms over users and threads

At its core, this section engages three factors or measures internal to the MZEE corpus that, when calculated for a word $w$ in an earlier window of time $T_i$, are theorized to predict the change in frequency of $w$ in a later timespan $T_{[i:i+1]}$ as in the previous section. These measures are: log-frequency of $w$ at $T_i$, dissemination over users $D^U$ of $w$ at $T_i$, and dissemination over threads $D^T$ of $w$ at $T_i$. The analysis presented takes as its model the system described in Altmann et al. (2011), adjusted for application to the present data.

6.5.1 Dissemination and dispersion in previous research

Dispersion has long been a consideration in corpus research; Gries (2010: 280) notes that “[t]here is too large a number of dispersion measures [...] to discuss them all, and unfortunately there is also as yet little work, let alone agreement, on which measure is best.” Dispersion measures, in basic terms, represent a word's distribution in a text, generally estimated by dividing a text into chunks of equal size and determining the number or percentage of chunks the word appears in. Such a dispersion measure is used by Chesley & Baayen (2010) as one of the predictors in their model of loanword frequency in French newspaper corpora. These sorts of dispersion measures, however, often strongly correlate with frequency, a problem for statistical models which assume independent predictor variables. To solve this, Chesley & Baayen (2010: 1355) regressed log frequency on log dispersion, retaining the model's residuals as their

---

50 I maintain the convention here of indicating generalized discussion of time windows/spans using a capital $T_i$; lower case $t_i$ is used to label specific time windows in this section.
frequency variable, so that frequency in that study could be considered that portion of frequency information unaccounted for by dispersion. More complex measures of a word's dispersion can take into account different ways of chunking a text, and be designed in a way so that the dispersion measure accounts for frequency and is therefore theoretically independent of that measure. Dissemination over users and dissemination over threads are two such measures which are particularly well-suited for forum data.

Taking as their baseline model an approximation of a Poisson process, Altmann et al. (2011: 1) introduce two measures of dissemination, $D^U$ and $D^T$, which are effectively operationalizations of the concept of word niche, defined as “the relationship between the word and the characteristic features of the environments in which it is used.” They found that $D^U$ and $D^T$ were both better predictors than log frequency in a given $T_1$ of an increase or decrease in log frequency in a given $T_2$ (i.e., the ultimate success or failure of a word) in their study of all words in two corpora drawn from Usenet discussion groups. The present analysis seeks first to determine whether these measures, applied to anglicisms, function as reliable determinants of anglicism frequency change, and second, to investigate differences in these factors' contribution to frequency change in anglicisms vs. native German words, i.e., to determine whether anglicisms pattern differently from German words in terms of their sensitivity to frequency and dissemination.

6.5.2 $D^U$ and $D^T$

Dissemination over users and threads, or $D^U$ and $D^T$, are calculated by first estimating for each word the result of a Poisson process, a random process that involves events occurring in a non-deterministic manner within a certain span of time. In a text corpus, this is analogous to shuffling all words in the corpus, holding constant the number of occurrences of word $w$, the number of users/threads, and the corpus size, with each word then being produced at random. The equation below is an estimation of this process: $U_i$ is the probability that a given user would use word $w$ at least once in a randomly-shuffled corpus, given that said word has a number of occurrences $N_w$.

---

51 A Poisson process is defined as:
“A process satisfying the following properties: 1. The numbers of changes in nonoverlapping intervals are independent for all intervals. 2. The probability of exactly one change in a sufficiently small interval $h \equiv 1/n$ is
$P = v h \equiv v/n$, where $v$ is the probability of one change and $n$ is the number of trials. 3. The probability of two or more changes in a sufficiently small interval $h$ is essentially 0.” (Weisstein 2012).
In (1a), \( m_i \) is the number of words that user \( i \) contributed to the corpus, and \( N_A \) is the number of words in the corpus. In essence, this equation, working from the innermost set of parentheses, considers the corpus size less \( j \), a counter that keeps track of the number of instances of \( w \) which have been stepped through. The size of user \( i \)'s contribution to the corpus, \( m_i \), is divided by this value: this yields the probability that the next occurrence of \( w \) in the corpus will be uttered by user \( i \). The complement of this fraction is taken by subtracting it from 1, yielding the probability that the next occurrence of \( w \) in the corpus is not uttered by user \( i \). These inverse probabilities are multiplied together for all occurrences of word \( w \), yielding the probability that user \( i \) did not use word \( w \) at least once. The complement of this, then (calculated by subtracting that probability from 1) yields \( \tilde{U}_i \), the probability that user \( i \) used word \( w \) at least once.\(^{52}\)

\[
(1a) \quad \tilde{U}_i = 1 - \prod_{j=0}^{N_w-1} \left( 1 - \frac{m_i}{(N_A - j)} \right)
\]

Once \( \tilde{U}_i \) has been calculated for all users for word \( w \), the probabilities are summed to produce \( \tilde{U} \), the expected number of users of word \( w \) under the assumption of this baseline model:

\[
(1b) \quad \tilde{U} = \sum_{i=1}^{N_U} \tilde{U}_i
\]

Finally, dissemination over users for word \( w \) is expressed as a ratio of the actual number of \( w \)'s users \( U_w \) to the estimated number of users of \( w \), expressed by the term \( \tilde{U} \). Thus, a word which is disseminated over users exactly as predicted by the baseline model would have a \( D^U \) of 1, while a \( D^U \) of less than 1 would indicate under-dissemination (a word used by a smaller number of users than expected given its frequency) and a \( D^U \) greater than 1 would indicate over-dissemination (e.g., a word used by a larger number of users than expected given its frequency):

\[
(1c) \quad D^U = \frac{U_w}{\tilde{U}}
\]

The calculation of \( D^T \) is analogous in every fashion, but substituting threads (and the symbol \( T \)) for users (and the symbol \( U \)). Note also that \( m_i \) is no longer the user's entire contribution to the corpus, but rather the number of words in thread \( i \):

\(^{52}\) (1a-c), while taken from Altmann et al. (2011) and equivalent to the original equations, have been expressed in a slightly different manner than in the original article in order to facilitate explanation.
While a $D^U$ or $D^T$ value of 1 is taken to be a word that is disseminated 'as expected' in a corpus, almost all words (including function words) have a tendency in real life towards burstiness or clumping. For this reason, it is expected that these dissemination values for words in the corpus will be, in general, less than 1. The final predictor, word frequency in the earlier time window $T_i$, is calculated in the usual way as a relative frequency: by taking the number of occurrences of word $w$ during the window and dividing it by the total number of words that occur in $T_i$. The relative frequency is then log-adjusted to yield $\log_{10} f$.

The final element is the response variable, which as in the previous analysis is expressed as $\Delta \log_{10} f$, i.e., the change in $T_{[i:i+n]}$ in the logarithmic value of the frequency of $w$—whether the word in question occurs relatively more or less often in the second time window $T_{i+n}$ than in $T_i$. In this case, a negative value for $\Delta \log_{10} f$ indicates a decrease in relative frequency, and a positive value indicates an increase in relative frequency.

To summarize up to this point, the dissemination measures $D^U$ and $D^T$ and relative frequency $f$ in time window $T_i$ are here being independently tested as predictors, with $\Delta \log_{10} f$ from an initial time to a later time as the quantities to be predicted. Prediction in this case, as instantiated in Altmann et al., is gauged by the squared correlation coefficient of each predictor with the response variable—the coefficient of determination $R^2$. 

\begin{align*}
(2a) \quad \tilde{T}_i &= 1 - \prod_{j=0}^{N_w-1} \left(1 - m_i / (N_A - j) \right) \\
(2b) \quad \tilde{T} &= \sum_{i=1}^{N_T} \tilde{T}_i \\
(2c) \quad D^T &= T_W / \tilde{T}
\end{align*}
6.5.3 Results

In contrast to the year-long time windows used in the Covo-MZEE analysis in section 6.4, the within-corpus analysis of predictors in MZEE uses six consecutive, non-overlapping time windows of 2 million words in size from the corpus, labeled as $t_1$ through $t_6$. This means that while the time windows do not strictly correspond to equal lengths of real-world time, the effects of differential subcorpus size are avoided: the likelihood of finding an (especially low-frequency) word in a subcorpus of 3 million words is, after all, greater than finding a word in a subcorpus of 1 million words, even when relativizing frequency. As in the previous analysis, predictor variables (this time including the dissemination measures) were measured at an initial time window $T_i$, and the response variable $\Delta \log_{10}f$ was measured in a timespan of length $l T_{[i : i+l]}$. As before, predictors and the response variable were aggregated across timespans of equal length, so that values from predictors at $t_1$ and $\Delta \log_{10}f$ from $t_{[1:3]}$ were aggregated with values from predictors at $t_2$ and $\Delta \log_{10}f$ at $t_{[2:4]}$, and so forth. This aggregation was done for predictor variable timespans of length 1 through 3. A gap of length 1 between the predictor variable time window and the beginning of the timespan for $\Delta \log_{10}f$ (e.g., predictors at $t_1$ and $\Delta \log_{10}f$ from $t_{[2:4]}$) was also tested, but those correlations were not found to be consistently significant. As in the cross-corpus analysis, correlations and coefficients of determination $R^2$ were calculated between the predictor variables ($\log_{10}f$, $D_U$, and $D_T$) and $\Delta \log_{10}f$. These correlations were performed for two sets of words: the set of all words in the corpus excepting the 850 known anglicism wordforms, and that set of anglicisms.\(^53\) The results of this analysis are reported in Table 6.3.

When measured over all words excluding anglicisms, $\log_{10}f$, $D_U$, and $D_T$ at an initial time $T_i$ are very weakly ($0.0309 < R < 0.0692$), but significantly ($p < .0001$) positively correlated with $\Delta \log_{10}f$ (left-hand side of Table 6.3). Considering the coefficients of determination $R^2$ for these comparisons, however, it is clear that there is a strong predictive relationship between these predictors and $\Delta \log_{10}f$, i.e., the best predictor is frequency at timespan length 1, and even that variable explains less than half of a percent of the variance in $\Delta \log_{10}f$. Moving to the right half of Table 6.3, the results for $D_U$ and $D_T$ over the set of anglicisms are significant ($p < .0000001$) but

\(^{53}\) An earlier analysis comparing the predictors for a set of 327 native German words frequency-matched to 327 of the anglicisms showed that the frequency-matched German words behaved like the (much larger) set of all non-anglicism words in regard to prediction of frequency change by $\log_{10}f$, $D_U$, and $D_T$.  

118
Similarly unimpressive in terms of prediction strength. Frequency, however, breaks away from $D^U$ and $D^T$ as a predictor, with a much higher coefficient of determination which increases along with the length of the timespan under consideration.$^\text{54}$

<table>
<thead>
<tr>
<th>Non-anglicism words</th>
<th>R</th>
<th>p</th>
<th>$R^2$</th>
<th>N</th>
<th>Anglicism wordforms</th>
<th>R</th>
<th>p</th>
<th>$R^2$</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{[i:i+1]}$</td>
<td>0.0566</td>
<td>&lt; 0.0001</td>
<td>0.32%</td>
<td>105553</td>
<td>$D^U$</td>
<td>-0.0814</td>
<td>&lt; 0.000001</td>
<td>0.7%</td>
<td>4145</td>
</tr>
<tr>
<td>$T_{[i:i+1]}$</td>
<td>0.0618</td>
<td>&lt; 0.0001</td>
<td>0.38%</td>
<td>105553</td>
<td>$D^T$</td>
<td>-0.0877</td>
<td>&lt; 0.000001</td>
<td>0.8%</td>
<td>4145</td>
</tr>
<tr>
<td>$T_{[i:i+2]}$</td>
<td>0.0505</td>
<td>&lt; 0.0001</td>
<td>0.26%</td>
<td>82223</td>
<td>$D^U$</td>
<td>-0.1207</td>
<td>&lt; 0.000001</td>
<td>1.5%</td>
<td>3306</td>
</tr>
<tr>
<td>$T_{[i:i+2]}$</td>
<td>0.0570</td>
<td>&lt; 0.0001</td>
<td>0.33%</td>
<td>82223</td>
<td>$D^T$</td>
<td>-0.1373</td>
<td>&lt; 0.000001</td>
<td>1.9%</td>
<td>3306</td>
</tr>
<tr>
<td>$T_{[i:i+3]}$</td>
<td>0.0470</td>
<td>&lt; 0.0001</td>
<td>0.22%</td>
<td>61444</td>
<td>$D^U$</td>
<td>-0.1634</td>
<td>&lt; 0.000001</td>
<td>2.7%</td>
<td>2471</td>
</tr>
<tr>
<td>$T_{[i:i+3]}$</td>
<td>0.0549</td>
<td>&lt; 0.0001</td>
<td>0.30%</td>
<td>61444</td>
<td>$D^T$</td>
<td>-0.1755</td>
<td>&lt; 0.000001</td>
<td>3.1%</td>
<td>2471</td>
</tr>
</tbody>
</table>

Table 6.3. Correlation results for predictors on non-anglicisms/anglicisms, timespan length 1-3.

This result contrasts sharply with the findings of Altmann et al. (2011), who found $D^U$ and $D^T$ to be more strongly predictive of frequency change over all words in their Usenet corpora: the results of the present analysis, then, suggest that anglicisms are a special case, and that they are sensitive to these factors in a fundamentally different way than are non-anglicisms. A second, and perhaps prima facie counterintuitive, finding here is that the correlation coefficient $r$ between frequency at an initial time and frequency change for anglicisms has a negative value, indicating that high initial frequency for anglicisms predicts a decline in frequency, and vice versa. To reiterate, popular anglicisms tend to lose frequency, while lower-frequency anglicisms tend to increase in frequency. At first, this sounds like a case of regression toward the mean, but if it is that sort of statistical artifact, a similarly strong effect should be seen with the non-anglicisms. Essentially, it seems that anglicisms are faddish and/or strongly connected to ingroup identity (a function that is no longer fulfilled once they are used frequently or spread

---

54 A parallel analysis with the data for timespans of length 1 and 2 truncated to the $N$ of timespan length 3 was performed; this increase with length of timespan was still evident, i.e., it is not an artifact of the varying $N$.
beyond the group, causing their eventual decline). This possibility is explored further in the conclusions to this chapter.

6.5.4 Discussion of analytical issues

Scatterplots of frequency, $D^U$, and $D^T$ versus the response variable for non-anglicism words and anglicisms are shown in Figures 6.6-6.7.

**Figure 6.6.** Scatterplot of initial log-frequency vs. log-frequency change, timespan 3.

**Figure 6.7.** Scatterplots of $D^U/D^T$ vs. log-frequency change, timespan 3.
In Figure 6.6, it is evident that there exist two floor effects for log-frequency change. This is due to the decision to use log frequency (rather than absolute/relativized frequency—relativization of frequency to corpus size is built into this analysis, as \(t_1-t_6\) are of equal size in words). As mentioned earlier, this is a standard design choice for corpus analysis. As a result, it was impossible to include words with a frequency of zero at either the predictor time or the time window ending the timespan.\(^{55}\) In addition, because of the particularly long tail of hapax and dislegomena in the corpus, it was mandated that words included in the prediction/response comparisons occur at least 3 times in the initial time window \(T_i\) (and at least once in the final time window \(T_{i+1}\)). For this reason, there are no data where \(\log_{10} f < 0.4771\). The second floor effect, seen in the lower left of Figure 6.6, is due to the logical maximum decline for low-initial-frequency items (negative frequency change for a word with initial \(f\) of 5 is capped at 4). This floor effect in particular might lead to an artificial negative correlation.\(^{56}\) Unlike the set of all words, though (where no such negative correlation was found), the set of anglicisms does not seem to cluster near enough to this boundary to be particularly affected (remember that these anglicisms are of overall corpus frequency > 100). The primary way to avoid this second floor effect is to increase the effect of the first; mandating that words have \(f > 50\) in \(T_i\) seems to avoid most of this effect (raising the other floor effect to \(\sim 1.69\) on the \(x\)-axis), but this has the unfortunate effect of removing most of the data, leaving only the most-established words for analysis. Because of the distance of the primary cluster of anglicisms from this boundary—and the fact that the negative correlation was found only for this set of 850 words—I suggest that it is highly unlikely that the anglicism findings were compromised by this effect.

6.6 Conclusions and implications

In this chapter, corpus statistics, including (limited) demographic information and the chronological shape of the corpus, were presented. In addition, the prevalence of anglicisms, and the increasing relative frequency of known anglicism forms were demonstrated. In two related analyses, I 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 of the English and German hip hop

\(^{55}\) \(\log_{10} 0\), for this purpose, can be considered an undefined value.

\(^{56}\) Thanks go to Janet Pierrehumbert for helpful notes on this phenomenon.
communities, and the relationship between the two, demonstrating that English frequency correlates positively with change in a borrowed word’s frequency in the German community—this result is not particularly shocking, and it is crucial to note that I do not posit a direct causal relationship between frequency in the Covo forum and frequency change in MZEE, but only an indirect one because the communities are exposed to many of the same primary texts (e.g., hip hop lyrics and material). It is, however, interesting that 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 might affect a word’s success or failure.

The second analysis, which compared three measures used by Altmann et al. (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. As evidenced in Fig. 6.8, where Altmann et al.’s findings with regard to coefficients of determination for the predictor variables are directly compared with the findings for anglicisms in this study, the reverse was found to be true of all words in their Usenet corpora.

![Figure 6.8. Comparison of \( R^2 \) values measuring predictive ability of variables in the present study (left) and in Altmann et al. (2011).](image)
In addition, the strength of the predictive power of frequency increases as longer time spans are considered, though this correlation is inverse—words with high initial frequency tend to see a decrease in frequency over time and those with low initial frequency tend to see an increase, a surprising finding in light of the logical hypothesis that more frequent words are likely to be seen by other users and hence more readily available for (re)production in a community. This finding, however, is not entirely unprecedented. In their diachronic study of loanword frequencies in two French newspaper corpora, Chesley & Baayen (2010: 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.” (dispersion alone having a net positive effect in their model).

A view of language as a stylistic resource (see, e.g., Coupland 2007) provides some explanation for these initially counter-intuitive findings: An anglicism which is used less often early on in a corpus, but which nonetheless survives, is likely to increase in frequency as other speakers adopt it in order to profit from its exclusive prestige within the group. However, a highly frequent anglicism seems to become increasingly undesirable—after all, if everyone is using a word, it may lose its capacity to distinguish in-group members (consider, e.g., the widespread adoption of the term *bling* outside hip hop culture in the US). Under this explanation, this circumstance is reflected by a drop in frequency as the word becomes passé and falls from fashion. However, if this is the case, there remains the possible argument that this is then not a phenomenon related to the status of particular words as anglicisms, but rather to their status as slang or in-group language—in which case, similar findings would be expected for German-language hip hop slang like *derb* 'rough', *alter/digga* 'dude' and others mentioned in the introduction to this chapter. Unfortunately, testing this hypothesis is made difficult by the lack of a sizable list of German-derived slang—making this a desirable task for future research. In any case, the anglicisms' status as in-group slang is not entirely divorced from their status as English-derived forms; in the ethnographic interviews reviewed in the next chapter, Chapter 7, the perceived properties of, and ideologies surrounding, English make it especially suited for the contribution of youth/in-group language to German, and the relation of those findings to the findings in this chapter form a major component of the concluding chapter of this dissertation, Chapter 8.
In summary, the analyses discussed here leverage the opportunities provided by large-scale corpus analysis of word dynamics and by the centrality of language to practices in 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.
CHAPTER 7  
THE ROLE OF IDEOLOGY IN LINGUISTIC PRODUCTION AND STYLIZATION

Ich muss ständig einsam fighten gegen Regeln, die ich nicht versteh' und dann enttäuscht begreifen, dass die Erde sich ja trotzdem dreht. 

[...] 

Möchte Kollegen bewegen, noch mehr zu sein als Strategen. 
Ey, ich hab' nix und bin niemand, doch rappen heißt für mich reden.

I always have to fight alone, against rules that I don't understand and then realize, disappointed, that the earth keeps turning anyhow. 

[...] 

I want to move colleagues to be even more than strategists Hey, I have nothing and I'm nobody, but to me, rapping is speaking. 

– Pyranja on 'E.H.M.' by Roey Marquis II, also feat. Schivv & Tatwaffe, 2002

7.1 Introduction and background

The previous three chapters have discussed issues centered on the how of contact-induced change, including methods of identifying anglicisms, a number of particular morphological forms that seemingly run counter to earlier observations about nativization, and the prediction of anglicism fate. What has not yet been addressed in an empirical manner is speakers’ motivations behind the use of anglicisms—and in an analysis concentrating only on linguistic forms, proposed explanations can only be speculative or provisional. For this reason, it is precisely these motivations that will be addressed in this chapter. These motivations find their fullest expression in the form of linguistic attitudes toward English, toward German, and toward anglicisms in particular by members of the hip hop community. These elements will be examined here using ethnographic data collected for the purpose. In addition, this chapter will tie these language attitudes to language ideologies, exploring the landscape of linguistic ideology in Germany and elucidating the process of production and reproduction of language ideologies within and outside the hip hop community.

This chapter, then, will engage primarily with the concepts of language attitude and ideology. These constructs will be explored through the examination of previous work on language attitudes and language ideologies in section 7.2, culminating in the express statement of the ideology-attitude relationship in 7.3. In section 7.4 I present an overview of the ethnographic data used for the analysis of attitudes in the hip hop community, including interview data collected in Hamburg in the summer of 2010 and selected threads from the MZEE forum.
discussing the use of anglicisms in the hip hop community. Next, I discuss the transcription of interview responses in 7.5. In 7.6, I present a number of excerpts from off- and online ethnographic work, linking the revealed attitudes to the ideologies which circulate in the German sociolinguistic sphere. Finally, I discuss the revealed ideologies and their implications for anglicisms in the final section, 7.7.

In the analyses which follow, I investigate the following research question: which linguistic ideologies are circulating and operating in the German hip hop community, and how do they condition language use? More specifically, do hip hop fans, as part of a global culture of opposition sharing a collection of 'resistance vernaculars', resist the top-down operation of the standard language ideology complex (described in section 7.2.2) in favor of language ideologies specific to their group?

7.2 Language ideology

While the theoretical construct of language ideology is commonly used by researchers seeking to understand the relationship between speakers and language, the term has seen several definitions in the course of its use. In addition to this variability in definition, the scope of the term's use is often unclear.

7.2.1 Language ideology: considerations of definition

Kroskrity (2004: 497-498) provides a relatively recent and comprehensive review of a number of definitions of language ideology, including that of Rumsey (1990: 346), as “shared bodies of commonsense notions about the nature of language in the world,” which actually echoes Silverstein's (1979: 193) earlier definition of language ideologies as “sets of beliefs about language articulated by users as a rationalization or justification of perceived language structure and use.” Kroskrity also discusses a definition from Irvine (1989: 255), who conceives of language ideology as “the cultural system of ideas about social and linguistic relationships, together with their loading of moral and political interests.” Kroskrity (2004: 498) reaches the summary conclusion that language ideologies are “beliefs, or feelings, about languages as used in their social worlds.”
While these four definitions (and others too numerous to reproduce here) share a number of components (all conceive of ideology as a set of beliefs/ideas), they also vary greatly in several important regards. First, it is important to consider whether language ideology is more profitably conceived of as single or multiple in nature: is there but one language ideology per community? Per individual? A second consideration, related to the first issue, is whether language ideology(ies) is (are) a construct that operates on the individual level (as Kroskrity's summary definition seems to suggest) or whether language ideology is instead a group endeavor, or whether both are true to some extent. Finally, can ideology target or have as its object particular linguistic features, or does ideology instead operate on the level of language varieties?

As to the first question, I follow Kroskrity (2004: 503) (and the majority of others) in agreeing that:

\[...\] language ideologies are profitably conceived as multiple because of the plurality of meaningful social divisions (class, gender, clan, elites, generations, and so on) within sociocultural groups that have the potential to produce divergent perspectives expressed as indices of group membership. Language ideologies are thus grounded in social experience which is never uniformly distributed throughout polities of any scale. [emphasis Kroskrity's]

If a researcher wishes to acknowledge that there do exist meaningful social distinctions between groups of people, as most sociolinguists tend to, and if language ideology is linked (as, e.g., Irvine 1989 has it) to these cultural or social groupings, then it is reasonable to conclude that language ideology varies as well, or rather that there exist a multiplicity of ideologies. Because ideology is an abstract concept, the difficulty that is encountered in drawing lines between, for instance, language varieties, is reproduced here. Groups of any size are without exception heterogeneous, so it is important to remember that, like languages themselves, language ideologies are convenient, constructed abstractions over an internally diverse population of instances. In the case of languages, these instances are idiolects or even utterances. In the case of ideologies, I argue that the relevant instances are, in essence, language attitudes; when these attitudes are consistent, shared, and enduring, they give rise to language ideologies as an emergent phenomenon.
Kroskrity's suggestion regarding the plurality of language ideologies does not stand alone. This trend in researchers' views about ideology finds similar expression in an earlier summary article by Woolard & Schieffelin (1994: 71):

The new direction in research on linguistic ideology has also moved away from seeing ideology as a homogeneous cultural template, now treating it as a process involving struggles among multiple conceptualizations and demanding the recognition of variation and contestation within a community as well as contradictions within individuals.

While recognizing the heterogeneity and multiplicity of ideologies, this excerpt also touches on a related issue: can ideologies be individual? Taking up the question of scale (whether language ideologies are held at the group or individual level, or both), Blommaert (2005: 158) sketches the logical extremes:

[…] there are, on the one hand, authors who define ideology as a specific set of symbolic representations – discourses, terms, arguments, images, stereotypes – serving a specific purpose, and operated by specific groups or actors, recognisable precisely by their usage of such ideologies. On the other hand, there are authors who would define ideology as a general phenomenon characterising the totality of a particular social or political system, and operated by every member or actor in that system. [emphasis Blommaert's]

Blommaert is discussing ideology here in a broader, interdisciplinary sense, including political-cultural conceptions of ideology, but this distinction holds for language ideologies as well. It seems that ideologies can be conceived of on the grand scale as societal processes, and also on a smaller scale as group or individual processes. This is inherently problematic because, discarding for the moment the possibility that ideologies are homogeneously distributed through an entire society, the concept of an ideology which is something held only by a single person inherently conflicts with the general understanding of ideologies (see definitions above) as 'shared', 'social', and 'cultural'. It is then worth stipulating that anything called an 'ideology' must necessarily be shared by a group of some size—individually-held beliefs that are neither shared nor enduring are instead better considered attitudes.

Engaging the final question here, that of the target or object of ideologies, there seems to exist a tacit consensus that ideologies are operant on the level of the language, dialect, or to use a less problematic term, language variety. Reviewing the definitions discussed earlier, it is clear that ideologies may have some role in informing prescriptions on the use of particular linguistic features, but these are not constitutive or central to the notion of language ideology. I suggest that
instead, language ideologies engage with language varieties—borrowing strongly from previous definitions, then, I suggest a working definition of language ideologies as systematic beliefs about language varieties which are shared by groups of speakers.

7.2.2 Dominant language ideologies in Germany

There are several language ideologies which have been identified by researchers as dominant in the German sphere, prominent among which are Herderian ideology, standard language ideology, and language purism. These three in particular are closely interwoven and often held in tandem, and I will discuss the relationship between them here, as well as the issue of whether 'alternative' language ideologies might also be said to exist.

The ideology of the standard language has, as its most commonly-cited definition:

a bias toward an abstracted, idealized, homogeneous spoken language which is imposed and maintained by dominant bloc institutions and which names as its model the written language, but which is drawn primarily from the speech of the upper, middle class.

(Lippi-Green 1997: 64)

This is corroborated in the German case by Spitzmüller (2007: 265), who notes that the concept of an idealized standard German was promoted by the “educated bourgeoisie (Bildungsbürgertum)” as early as the 17th century, serving as a status symbol for this new and newly powerful class. Standard language ideology, at its core, is a system of beliefs that establishes the abstract concept of a standard, homogeneous language, treating it as real. This concept is echoed in Preston's 'folk theory' of language, drawn from interviews, which conceives of a 'real' language external to the individual speaker. This conception of a standard language underlies and serves as a prerequisite for the remaining two ideologies discussed here, the first of which is Herderian ideology.

Woolard (1998: 16-17) gives a concise overview of the history and content of what has been called Herderian ideology:
It is a truism that the equation of language and nation is not a natural fact but rather a historical, ideological construct. This construction is conventionally dated to late eighteenth-century German Romanticism and Johann Herder's famous characterization of language as the genius of a people, and thus it is often referred to as the Romantic or Herderian concept of language. [...] this Herderian or nationalist ideology of language is globally hegemonic today.

In effect, this ideology, which equates a single nation (with a single people) with a single language, is an interaction of the standard language ideology with the political ideology of the nation-state, and it is prevalent in modern discourses about German. Spitzmüller (2007: 270), in a corpus study of German media discourses about language (from the 1990s-early 2000s), finds that:

[...] for most discourse participants, it is taken for granted that there is a natural link between a particular national language and 'a particular' national identity (this is explicitly stated in 126 documents, whereas it is not questioned in a single document in the corpus).

Given the problematic (and constructed) nature of the nation-state (see, e.g., Hoffman 1966, Wimmer & Schiller 2002, and no small number of critiques in between), this lack of dissent is—from a critical standpoint—distressing. Do there exist ideologies which compete in the public sphere with Herderian and standard language ideologies? Before addressing this question in the next section, I discuss another identified ideology which, again, has its foundation in the standard-language and Herderian ideologies.

Spitzmüller (2007) provides an overview of language-ideological history in Germany through the German tradition of Linguistische Diskursanalyse, which takes modern discourses to be necessarily linked to the context of historical discourse. In engaging with the notion of language purism, Spitzmüller (2007: 261) finds that, like standardization, it has a complex history within the German language dating back to the 17th century. In the 1990s, after re-unification, language purism re-emerged as a hot topic, and has been prominent in media discourses since that time. According to Spitzmüller, this latest wave of language purism, expressed through anti-anglicism sentiment, reached a peak in 2000-01, with repeated calls for political action in the form of legal protection/enshrinement of German in the constitution as a national language. In a statement which further supports the identification of anglicisms as the primary concern of language purism, Stein (2005: 190) writes:
Purism has mostly been seen in terms of resistance to another language […]. Language mixing – arguably the most natural of all processes in which languages are involved – is seen as the major infringement of purity, both on the level of folklore beliefs about language and the level of scientific inquiry.

This ideology, then, takes a natural process—namely contact-induced language change—and recasts it as problematic. Woolard & Schieffelin (1994: 64) elaborate further: “purist doctrines of linguistic correctness close off non-native sources of innovation, but usually selectively, targeting only languages construed as threats.” The ideology of language purism, that is, does not treat all contact-induced change equally, but primarily disfavors change originating from certain language varieties—generally those seen either as foreign standard languages having an overbearing influence (as is the case of English with regard to German) or those not construed as proper languages (dialects, minority/immigrant languages) or languages of low prestige (see, e.g., Zentella (1997) for a related treatment of Spanish code-switching in the US). In addition, it is clear that language purism relies on the standard language ideology, as only a homogenized language can be threatened by change, outside influence, and heterogeneity. Furthermore, language purism is often (but not always) connected with Herderian ideology—the interaction of the nation-state ideology and standard-language ideology connects linguistic purity to national purity. In this way, Herderian ideology, the nation-state ideology, and standard-language ideology do not function independently of each other in the German sphere, but rather as an interacting ideological complex, which I refer to as the standard language ideology complex.57 I provide several illuminating examples of expressions of these ideologies, demonstrating their interaction in recent German media, in the next section.

7.2.3 Metalinguistic discourse in recent media
Spitzmüller (2007) notes that metalinguistic discourse about anglicisms in German media continued through 2001; in this section, I discuss more recent (post-2001) instantiations of the metalinguistic media debates over anglicisms, which provide a crucial (and contemporary) context to the ethnographic interviews that constitute the primary data in this chapter. Over the course of the research project, I collected discourses about language, and especially about

57 My use of this term should be understood throughout this chapter to apply to an ideological complex operating among speakers of German; I do not claim that these ideologies interact in the same way in other linguistic communities with standard language ideology.
anglicisms, in the German media in order to ground and contextualize my discussions with interviewees. Although I have not collected a large corpus of several hundred instances, as did Spitzmüller (2007), I present several more recent examples, to be taken in parallel with Spitzmüller's findings, which connect media discourses about anglicisms to language purism, Herderian ideology, and standard language ideology in a historical context. The excerpts presented here demonstrate that the prominence of anglicisms in German media discourse has continued through at least 2010.

In his analysis of language-ideological discourse in Germany, Spitzmüller (2007) analyzes the positioning of what he calls ideology brokers (after Blommaert 1999), specifically including published statements from purists, linguists, journalists, and politicians. In this section, I establish the continued relevance of anglicisms in German media by reproducing and discussing mediated metalinguistic discourses. I begin with a journalistic article on the issue from well-regarded national newspaper Die Zeit, titled Ist Deutsch noch zu retten?, 'Can German still be saved?' featuring a tagline reading Englisch ist die Weltsprache. Aber wir können verhindern, dass unsere Muttersprache weiter erodiert, 'English is the world language. But we can prevent our mother tongue from eroding further.' This article by Ulrich Greiner (see Figure 7.1) was prompted by renewed calls in 2010 for the establishment of German as an official national language in the Grundgesetz, or constitution. The article raises concerns about the future of German, citing a number of German linguists in arguing that English is becoming dominant not only globally but particularly in Germany, and making the case not for language purism per se, but rather for the restoration of German as an economic, academic, and political language. In a discussion of anglicisms in particular, Greiner (2010) writes:

Die Frage lautet, ob Dieter E. Zimmer recht hat, der schon 1997 befürchtete (in seinem Buch Deutsch und anders), das Regelsystem des Deutschen werde durch die Invasion des Englischen aufgeweicht, bis das intuitive Verständnis dessen, was sprachlich richtig sei, gänzlich verschwinge. In jedem Fall führe es dazu, dass uns die Sprachzustände der Vergangenheit fremd oder unverständlich würden.

The question is whether Dieter E. Zimmer is correct—he feared as early as 1997 (in his book Deutsch und anders) that the German language's system of rules would be weakened by the invasion of English until even the intuitive understanding of spoken correctness might completely go away. In any case, it would lead to a situation where the language of the past would become alien or incomprehensible to us.
First, language attitudes are revealed by the author in the tagline: German is discussed as *unsere Muttersprache* 'our mother tongue'. This attitude can be considered to express Herderian ideology; the discourse positions both author and audience as Germans, to whom the German language belongs. Under this conception of language, English words and phrases remain innately English, and as such, they erode the German language. In the excerpt above, this concept of erosion is repeated, and the linguistic fears are framed with regard to 'correctness' and the severing of a link with the languages of the past.
Policy reactions to these concerns in Germany are relatively rare, and, in general, newsworthy when they occur. Peter Ramsauer, the German Bundesminister of Transportation, made headlines in early 2010 for battling against anglicisms in his ministry—getting rid of positions and agencies like Travel Management and Task Forces in favor of Reisestelle and Projektgruppen.

Ramsauer, quoted in an interview with the newsmagazine Stern, states (Schütz 2010):

_Englisch ist eine Weltsprache, die die Menschen verbindet, und das ist auch in Ordnung. Wir aber leben in Deutschland und sprechen unsere Muttersprache. Ich kenne kein Land der Erde, in dem man so respektlos mit der eigenen Sprache umgeht. Millionen Bürger fühlen sich ausgegrenzt, wenn uns Anglizismen inflationär und willkürlich überfluten. Diese Entwicklung mache ich im Bundesverkehrsministerium nicht mit, nur weil es modern oder "chic" ist. Wir haben für jeden Bereich unseres Lebens auch deutsche Vokabeln, wir müssen sie nur nutzen._

[my translation]

Here, Ramsauer describes English as a 'world language', but the equation of nation to language in his statement—'we live in Germany', 'we speak German'—suggests that Herderian ideology informed his language policy decision. At the same time, we see one of the (albeit mainstream) non-standard-language ideologies discussed below, that of elite bilingualism, emerging in this statement. English can index modernity in Germany (Piller 2001), and Ramsauer's metalinguistic discussion acknowledges this—and it is important to note that unlike other contemporary German political figures like Guido Westerwelle and Gunther Oettinger, who are often mocked for their poor command of English (Crolly 2010), Ramsauer has no notable deficiencies in this regard. Still, Ramsauer plays on fears that many, especially older, Germans will be 'fenced out' (ausgegrenzt) by anglicisms. This metaphor also suggests a Herderian unity of national and linguistic borders, as does Ramsauer's use of Invasion, 'invasion'. The use of überfluten, 'overflow' invokes the common metaphor of the adoption of English words as a flood or other water-based disaster (Spitzmüller 2007). Finally, Ramsauer's statement that “we have German vocabulary items for every aspect of our lives” suggests, essentially, that anglicisms are
generally deemed unnecessary and unwarranted, i.e., in the above statement Ramsauer completely discounts the need motivation (Hock & Joseph 1996) for borrowing.

Given the tone of the majority of statements from this interview excerpt, it would be fairly simple at this point to suggest that Ramsauer is reproducing the standard-language ideological complex in several forms, through Herderian ideology to language purism to the ideology of the standard language. However, one statement he makes complicates this view. In the same interview, the Stern interviewer asks, in a tongue-in-cheek way, if Ramsauer might need a Break for abchillen 'chilling out' after the interview. Ramsauer's response: Natürlich chillen meine Töchter auch mal, das ist normale Jugendsprache, 'Naturally, my daughters chill every now and again, that's normal youth language'. This statement betrays a nuanced operation of language ideologies behind Ramsauer's attitudes, a configuration in which some anglicisms are seen as Jugendsprache 'youth language' and, as such, non-problematic. This connection between anglicisms and youth language, crucially, resurfaces in the interviews discussed later in this chapter—although, perhaps surprisingly, with quite a bit more concern about the negative societal effects of anglicisms. This being said, the Ramsauer interview, along with the Zeit article above, demonstrate the survival of all three components of the German standard language ideological complex outlined in section 7.2.2 in recent media discourses.

7.2.4 Alternative ideologies

Having identified and related the three ideologies that constitute what I have called the German standard language ideological complex, it is relevant to ask whether there are competing alternative ideologies which are regularly expressed in the German sphere; as Eckert (2008: 467) writes, “Variability across communities is not limited to linguistic form but is present in the understanding of what that form means and ultimately in the ideologies that underlie language use.” In this section, I address this issue specifically for the German-speaking sphere.

One such alternative might be found in the valorization of English (elite) bilingualism or internationalism. Piller's (2001) discussion of language indexicality in German advertising is not framed as a primarily ideological approach, dealing as it does with voicing and identity
construction. However, some of her findings point to the implicit representation of English as a language with a particular role in conceptions or constructions of identity (2001: 155):

 [...] although German advertising may construct both identities of the national Self and of the national Other as multilingual, bilingualism in English and German is set up as the “natural” option for successful middle-class Germans, while other languages (e.g., Italian, Russian, or Spanish) are presented as languages of the cultural and national Other.

Piller goes on to mention that English bilingualism in German advertising in particular indexes internationalism, future orientation, success, sophistication, and fun in advertising, while the neighboring Continental national languages of French and Italian are relegated to portrayals of eroticism and delicious food, respectively. Only in two advertisements (out of a large corpus), one by a non-profit bicyclists' union, and another by the language purist organization Verein Deutsche Sprache 'Association for the German Language', English is used ironically to criticize the values above. With these two exceptions, the prevailing ideology of multilingualism, then, and particularly of English bilingualism, is in direct opposition to language purism, just as the ideology of multiculturalism is in direct opposition to integration/ethnic nationalism. Piller notes a broader juxtaposition of ideologies that reach beyond language in her conclusion (2001: 181):

Commercial advertising is but one of many discourses available in contemporary Germany that offer people ways to think about their identities and their social and societal relations. Political nationalist discourses, for instance, set up quite different ideals in which national belonging is foregrounded and internationalism is perceived as a threat to one's identity.

English in advertising is only one domain where this ideology of elite English bilingualism is found—Hilgendorf (2005), for example, finds an overwhelming shift toward English in German foreign language education.

It is also relevant to investigate whether hip hop itself, as a globalizing culture, carries with it any ideologies of language. Specifically, since hip hop is often theorized as a subculture of resistance to the mainstream, it is pertinent to ask whether the notion of resistance in this case also includes resistance to the standard language ideology complex expounded in the previous section. One of the most oft-cited statements about language and hip hop is summarized by Mitchell (2002: 41):

Potter [(1995)] sees African-American rap as a form of “resistance vernacular” which takes the minor language's variation and redefinition of the major language a step further and “deform[s] and reposition[s] the rules of “intelligibility” set up by the dominant language.”
It has been noted by a number of authors (including Mitchell himself) that this notion of rap, and hip hop language more generally, as a 'resistance vernacular' is perhaps even more valid in hip hop culture's global instantiations than it is for the commodified hip hop language in the US. Even so, Alim & Pennycook (2007: 90-91) contrast the views on language proffered by US hip hop artists— which valorize marginalized language—with “dominant ideologies of monolingualism and monoculturalism.” Citing an interview with the rapper Jubwa conducted by Alim in 2000, the authors demonstrate attitudes that may be indicative of an ideological view of 'Black language' as limitless and unrestricted, positively evaluating the notion of breaking/changing rules. Alim's (2004) own work has itself been central to *flippin the script* by subverting Herderian ideology into a relationship between a (Global) Hip Hop Nation and (Global) Hip Hop Nation Language(s). However, work like Alim's on attitudes toward language in American hip hop culture generally stops short of identifying constructs like Hip Hop Nation Language as part of an ideology or ideological complex, and it is furthermore unclear whether, even if such an ideology exists, it is reproduced in hip hop contexts in Germany. Later in this chapter, I begin addressing this gap in knowledge through the analysis of ethnographic interviews with hip hop fans.

7.3 The relation of language attitude to language ideology

If ideologies, like those discussed in the previous sections, are ideologies by virtue of being systematic, shared beliefs about language, it becomes important for the purposes of the present study to discuss how, exactly, these ideologies can be identified in interviews with individuals. In this section, I discuss the notion of the *language attitude*, and how it relates to language ideology; this has the end effect of suggesting methodological considerations and reducing the risk of confusing a one-off attitude about language use with a more consistent group ideology.

To this point, it has not been especially common to integrate language attitude and language ideology into a particular theoretical schema. In one of the few mentions of the two concepts together in previous literature, Woolard and Schieffelin (1994: 61-62) note:

> Although the extensive body of research on linguistic prestige and language attitudes grew up in a social psychological framework [...], the intrapersonal attitude can be recast as a socially-derived intellectualized or behavioral ideology (Bourdieu's habitus).
It is true that language attitude research in the vein of Preston's work (see, e.g., Preston 2003 for a summary) is a domain where social-psychological experimental (in contrast to ethnographic or discourse-analytic) methods and data are generally used. The idea of intrapersonal attitude as a behavioral ideology, however, is one I take issue with. Behavioral ideology is a notion originating in Vološinov's work, where it is defined as an “atmosphere of unsystematized and unfixed inner and outer speech which endows our every instance of behavior and action and our every 'conscious' state with meaning” (Vološinov 1973: 91). In that work, behavioral ideology is contrasted with social ideology, and I find the former somewhat compatible with what I would term the individual reproduction of group ideology. However, I believe it is important not to blur the lines between this individual reproduction of ideology, the notion of intrapersonal attitude, and the expressed attitude. I understand intrapersonal attitudes as those held internally by an individual, and not all of these are necessarily expressed (for a multitude of reasons). What differentiates these internally held attitudes from the individual reproduction of ideology, however, is the systematicity of ideology. While attitudes are, in essence, beliefs/ideas about things, and ideologies are defined as sets of these beliefs, not every attitude is ideological or part of a systematic whole.

Preston (2003: 62-65) discusses the notion of a 'folk theory of language' which emerges from language attitude research, generalizing from a large set of experiments to identify notions of 'correctness' and 'pleasantness' as relevant dimensions for respondents' characterizations of American dialects. His discussion reveals a folk conception of a 'real' language external to human beings, against which dialects and errors are judged. This, then, being a system of beliefs about language held by a group of speakers, is what I would consider a language ideology, one which is identified through social-psychological elicitation of individual language attitudes. Certainly, not all of the attitude expressions collected by Preston and other researchers fit this theory—the key is that a large body of them point to this conception of language.

It seems that the primary difference between language attitude and language ideology is one of scale: attitudes are individually held and individually expressed beliefs, and can be (but are not

58 In addition, the mention of Bourdieu's notion of habitus in connection with an intellectualized ideology is perhaps inappropriate, as habitus is 'em-bodied', and as such, necessarily also physically manifested (Thompson 1991).
always) momentary or fleeting; attitudes are also not necessarily systematic. I argue that it is useful to distinguish ideologies from attitudes on the one hand by limiting the use of the former term to designate enduring and systematic beliefs which are shared by groups of speakers. On the other hand, the target, or object, of language attitudes versus ideologies is also a point of difference. Language attitudes may target specific linguistic forms or entire language varieties; ideologies, on the other hand, have as their object only linguistic varieties.

Language attitudes are thus sometimes (but not always) expressions of more generally held language ideologies (or, more specifically, the reproduction of those language ideologies in the mind of the individual). Attitudes elicited from speakers, then, can in the aggregate be seen to represent or indicate which language ideologies are instantiated in those speakers, and at work in the group. These attitudes may be explicitly stated, or implicitly inferred from use. When Kroskrity (2004: 505) writes of “ideologies of practice that must be read from actual usage”, I propose that such inferences are mediated by the identification of implicit language attitudes, and the subsequent generalization of those attitudes to linguistic ideologies. The seeming circularity in this process is one of the primary reasons why I find an ideology-only approach problematic, and suggest the inclusion of language attitude in this process. As Blommaert (2005:164) writes: “We see [...] distinctions in Vološinov's work, when he separates 'established ideologies' from 'behavioural ideologies' (Vološinov 1973: 90-92), in which the former is both the crystallisation of and the motive for the latter.” I submit that language attitudes are the missing step in this process, forming a cyclical process illustrated in Figure 7.2.

An ideology, as an abstract concept, is disseminated through a variety of channels, and influences the individual, who comes to reproduce that ideology internally in a more or less intact way. The individual then harbors a number of internal attitudes, some of which are ideologically informed, and others which may not be systematic. Some of these attitudes are expressed by the individual, either directly through the statement of attitude or in a less explicit manner. These attitude expressions may additionally be mediated and reproduced—and when these attitude expressions aggregate from multiple individuals, exhibiting some systematicity and ability to endure, they can be said to jointly constitute the supra-individual ideology (in the top left-hand corner), where the cycle begins again.
I also take here the critical view that ideologies are integrally linked to structures of culture and power. In terms of language attitudes as instantiations of ideologies, this means that the attitudes expressed by the powerful are more likely to be indicative of more influential linguistic ideologies (and more likely to be reproduced and widely disseminated). Attitudes expressed by these ideology brokers are more likely to be influential and representative of ideologies.

It stands to reason, then, that language ideologies can be investigated in several ways. The most direct way to gain access to popular language ideologies circulating in a community is to assess instantiations of those ideologies in mass media outlets, which serve as “key arenas for the production and reproduction of language ideology” (Androutsopoulos 2010: 182). Such mediated instantiations are the focus of section 7.2.3, above. However, only the most official or widely held language ideologies can be expected to be expressed via these outlets. When examining alternative ideologies which are often operative on a smaller or even individual scale, and where it is expected that these ideologies may be even more numerous and conflicting, I
choose to glean information about operant ideologies from attitude expressions in posts from the MZEE forums, elicited reactions to these posts from interview data, and unelicited attitudes expressed in those interviews, the collection of which is detailed in section 7.4. In every case, the context of the attitude expression, as well as its corroboration (or lack of corroboration) by established ideologies or other expressed attitudes, is carefully considered before any claim is made that the attitude is indicative of language ideology.

7.4 Data and Methodology

In order to shed light on the issue of language ideologies in the German hip hop scene, I will use a variety of ethnographic (broadly understood) and discourse-analytic methods, including data from the MZEE corpus alongside a range of research activities I undertook in Hamburg, Germany in the summer of 2010 and continuing online through 2012. I consider the fieldwork component of this project including real-life interactions with social actors in the hip hop community in Hamburg to be crucial to the success of the overarching project in providing the researcher with an insider perspective on German hip hop culture through immersion in a lively 'scene' and through contact and interviews with rappers, DJs, students and fans, which was necessary for the development of a participant-observer perspective. In addition, the research trip provided an immersion in mainstream, everyday German culture by and large not available online. In this section, I will outline the types of data and data-collecting methods used for this part of the project.

First, while examining the MZEE forums throughout the computational analyses, I came across a number of threads containing fans' metalinguistic comments about anglicisms. These threads were catalogued and are used as one source of data in this chapter. Several posts from another thread on the topic of anglicisms (not included in the MZEE corpus, which contains only threads from the Hip-hop Diskussion subforum) were also used as interview prompts in the ethnographic interviews, which form the second data source analyzed here.\footnote{This was the thread 'Anglizismen!!!' from the off topic section of the MZEE forum, from which selected posts were used as interview prompts, [http://www.mzee.com/forum/archive/index.php/t-55094.html] Accessed 17 April 2012.}
In order to collect data firsthand in the German hip hop scene, I was able to spend the period from May 15 – August 15, 2010 in the city of Hamburg. Hamburg is Germany's second largest city (and a federal state in and of itself) and home to one of the earliest prominent hip hop movements in Germany, centered around the popular success in the late 1990s of, among others, the rap group Fettes Brot and the *Eimsbush* record label. *Eimsbush* is a portmanteau word drawing from Eimsbüttel, a neighborhood in Hamburg, and *Flatbush*, a neighborhood in Brooklyn, NY, known for producing a number of rappers, like Busta Rhymes, Talib Kweli, and Fu-Schnickens, in the 1990s. *Eimsbush* Entertainment, while now defunct, was founded by rapper/reggae/pop artist Jan Eißfeldt (a.k.a. Eißfeldt 65/Eizi Eiz/Jan Delay), and was home to nationally-known rappers and DJs, including Samy Deluxe (and his group Dynamite Deluxe), Illo, and D-Flame. *Eimsbush* artists also collaborated frequently with other artists in the area and nationally, releasing several mixtapes known as the *Style Liga* (Style-league) compilations and forming a late 1990s-early 2000s collective known as the *Mongo Clikke* (roughly, 'Retard Clique'). The influence of this period, and especially of the *Eimsbush* label, remains today: as of 2010, Hamburg's public trashcans were given comic-style speech bubbles containing colloquial and attention-getting sayings, including a trashcan I found in the Eimsbüttel neighborhood, pictured in Figure 7.3. Since the early 2000s, Hamburg's prominence in the German hip hop scene has receded somewhat, eclipsed by the massively popular but short-lived *Aggro Berlin* label and the ongoing metastasis of German hip hop to niche scenes and regional rap from smaller cities. Hamburg, however, is still home to a thriving scene. During the course of the research project in 2010, I was able to attend numerous hip hop-related events, including several DJ nights and freestyle battles, as well as a few smaller concerts from up-and-coming hip hop artists. In addition, some of my data comes from interactions with students and instructors at the Hip Hop Academy Hamburg, a non-profit youth program (for ages 13-20, per promotional materials) which is supported by the city-state of Hamburg.

---

60 I am grateful to the DAAD (German Academic Exchange Service) for the Graduate Research Grant which enabled this research visit and to Prof. Dr. Jannis Androutsopoulos who extended an invitation and kindly agreed to advise the research project during that visit.

61 Current information can be found at the Academy's website, [http://www.hiphopacademy-hamburg.de]. Accessed 12 March 2012.

62 My interviews were conducted solely with subjects over the age of 18, per the IRB agreement guidelines (see Appendix D).
The Academy offers what is essentially an after-school program as well as a more intensive summer program, featuring instruction from well-known Hamburg hip hop veterans. Courses offered in 2010 included rapping, DJing, breakdancing, graffiti art, 'Nu-Style' dancing (choreographed hip hop dance, contrasted with the more 'freestyle' esthetic of breakdancing), and beatboxing (the production of 'a capella' beats using vocal apparatus). Figure 7.4 shows an advertising postcard for the Hip Hop Academy, in which English material is prominently used. While in Hamburg, I made contacts in the hip hop community primarily through (offline and online) social networking in combination with direct contact. As an example, I began by contacting a Heidelberg-based radio DJ I was acquainted with, who then provided several contacts with his friends in Hamburg. One of these friends put me in contact with his brother, a

---

63 *Digga* is a verbal form iconic of Hamburg dialect probably (though folk-etymologies abound) from Ger. *Dicker*, 'fat/thick guy', and is used as a near-universal term of address, especially among young to middle-aged males in Hamburg. *Digga* (along with a more widely used form characteristic of Berlin, *Alter/Alda/Oida*) is also associated with, but not exclusive to, hip hop culture. One might speculate that the similarity in phonological and orthographic form to a Black English and Hip Hop Nation Language term of address has strengthened this association. Besides this lexical item, the phrase 'Keep your Eimsbush clean' is written in English, and along with the portmanteau *Eimsbush*, with its associations with American hip hop history, this language choice constructs the neighborhood as particularly oriented to American hip hop culture (although Eimsbüttel today is a relatively gentrified neighborhood). The native regional form *Digga*, on the other hand, localizes the discourse.
student at the Hip Hop Academy, who in turn introduced me to the program director of the HHA, who I then arranged to meet—both to interview him about the Academy's goals and to ask for permission to interview students and instructors.

Figure 7.4. Promotional postcard for the HHA Hamburg, featuring prominent use of English.

In order to interview a more representative sample of the German hip hop community, only a portion of the interviews were conducted at the HHA. A number of other contacts outside the Academy were also made through social networking, both through members of the hip hop community who I already knew and through interactions at hip hop events. Between these contacts, I was able to record 10 interviews with hip hop community members in Hamburg, as well as two sessions of natural conversation as a participant observer during producing and graffiti courses at the HHA. I was also able to take notes on two unrecorded interviews and collect stray observations. In all, five of the recorded interviews were with HHA students, one
was with an HHA instructor and local hip hop DJ, and four were with non-HHA-affiliated members of the hip hop community, including a Hamburg-based rapper and the radio DJ from Heidelberg. These interviews ranged from thirty minutes to three hours in length, depending on the individual interviewee's interest and time constraints. Interviews were recorded with a digital audio device, and structured according to a flexible interview protocol. The protocol evolved, in part through trial and error, throughout the interviewing process. In particular, the questions used in several early interviews were found to not be open-ended enough, and later interviews included additional material, as I came across advertisements and other relevant ephemera to use as interview prompts. The final interview protocol is reproduced in Appendix C. As noted earlier, all interviewees confirmed that they were 18 years of age or older at the time of the interview, but I did not always ask for interviewees' exact ages; a few of the hip hop instructors and artists interviewed, in particular, being veterans of the German hip hop scene since the 1980s, could be considered older than the average hip hop community member. Considering the youth-orientation of the subculture, it was determined that in these cases, asking for the exact age of the interviewee could color interview responses by making this issue salient. In general, my goal was to keep the interview (while clearly defined as an interview) as natural as possible. In the beginning of each interview, I read a script (approved by the University of Illinois campus IRB) which asked if I could record the interviews, clearly informing the interviewees that any publication of their responses would be anonymized, and that they were not compelled to answer particular questions and could withdraw from the study at any point. The oral consent script used at the beginning of each interview can be found in Appendix D. The IRB approval letter for this research is also included in Appendix E.

7.5 Transcription of spoken interview responses

In the analyses below, linguistic attitudes expressed by hip hop community members from the MZEE forums are presented alongside linguistic attitudes elicited during the ethnographic interviews performed in Hamburg. While the faithful reproduction of written forum data is largely unproblematic, speech transcription carries with it a number of problems. Bucholtz (2000, 2007) rightly notes that linguistic transcription involves interpretive and representational decisions on the part of the transcriber, corresponding to, respectively, the questions of what is transcribed, and how it is transcribed.
I am, in essence, trying to identify attitudes to language as they are expressed by interviewees and recognize therein the influence of established and emergent language ideologies. Because written transcriptions of speech are always a negotiation between what was spoken and what was understood by the researcher, I unreservedly stress here that the transcriptions of interviewee responses provided in this section reflect my understandings of those responses. As Bucholtz (2000: 1440) writes:

Ethnographies [...] are not transparent and unproblematic records of scientific research, but are instead creative and politicized documents in which the researcher as author is fully implicated.

My aim here is neither to analyze the organization or structure of conversation nor to assess phonetic detail. Instead, in my transcriptions, I seek to reproduce interviewees' responses as accurately as possible on the lexical level while adhering to general conventions established in the reproduction of speech for a non-specialist audience. In the interest of engaging in a sort of 'reflexive' analysis (Bucholtz 2000), I consider it important to note in this section some of the decisions I have made in transcribing and the reasoning behind them, additionally revealing in this way my own contribution to the reading of these transcribed interviews.

In the excerpts reproduced here, I adopt the convention of presenting interviewees' responses to questions as fairly complete and coherent statements, naturalized to many of the standards of lay reporting of speech.64 I note that there are particular pitfalls associated with this transcription style—pitfalls made extremely clear through Bucholtz's (2007) reanalysis of some of her earlier transcriptions. One in particular, which she initially presented as a sort of 'nerd manifesto', was presented as a clear, unhedged statement; the same data, when transcribed more closely, revealed a number of hedges, false starts, and interventions by the interviewer which call the confidence of the statement as originally transcribed (and its characterization as a manifesto) into serious question.

In order to avoid the issue of overstating the confidence of statements that I transcribe and reproduce here, I have chosen to include nonverbals (reproduced in German as, e.g., ähm or em, and in the English translations as um or er), hedges, and words with discourse-marking function

64 Thanks go to Sarah Henneböhl and Christoph Baumeister for their help in this process; remaining errors are of course my own.
such as German also 'so' which appear in the recordings. However, I generally transcribe and include my own utterances only when they are not functioning simply as continuers; interactional utterances like mhm and ja on my part are not, in general, reproduced here. Parts of the interview internal to an excerpt which are not reproduced are indicated by a bracketed ellipsis […] and incomprehensible or unintelligible portions with [unint.] Any other material in brackets is my addition to the transcript and included as a suggested clarification or other note. Prosody is largely unrepresented; emphasis and stress where particularly prominent are indicated with underlining. I also use the punctuation marks . and ? generally to indicate intonational contours of preceding material—a period indicates falling intonation, a question mark rising intonation. The comma is used in a less circumspect manner to indicate smaller intonation units, short pauses, and where necessary to improve readability of the transcript, and the hyphen - is used to indicate a discontinuity or 'cut-off'. Anglicisms which are not well-established that are found in the text are presented in bold face in the original and translation; this should not be read as emphasis (which is indicated, as noted above, with underlining) but merely as an aid in identifying the use of anglicisms in the discourse.

In the transcription of speech, the faithful representation of dialect is another common problem. Bucholtz (2000: 1454) summarizes Preston's arguments to this end:

Preston […] rejects the use of nonstandard orthography in academic writing to represent colloquial and vernacular speech. He argues that such spellings call up a negative image of the speaker from which even trained scholars may not be entirely immune. Instead of the unsystematic approach taken by researchers who employ unconventional orthography, Preston proposes a set of guidelines for the sociolinguistically informed transcription of folklore. These guidelines are summed up in Preston's declaration, “I believe that morphological accuracy is the appropriate level [of representation] and that phonetic precision should be sought only when that level is pertinent to the lore or the clarity. (Preston 1982: 323)”.

In deciding how best to transcribe spoken interview data, there is also the matter of my own proficiency in German; while I am conversationally near-native in a fairly deregionalized standard German (Hochdeutsch), my familiarity with other varieties of German (and any possible eye-dialect or 'standard' colloquial orthographies) is only moderate; a systematic attempt

---

65 Transcribed statements prefaced with M: are my conversational contributions, and as I am not a native speaker of German; a number of grammatical errors remain in those conversational turns.
on my part to reproduce various elisions common in informal speech (like $n$ for *ein/eine* or *haste* for *hast du*) would likely be stilted and inconsistent, and could additionally (as Preston notes in Bucholtz’s discussion above) prove problematic in that it could predispose readers toward negative or otherwise erroneous attitudes toward the interviewees, when the interviewees’ own language attitudes (and the underlying ideologies) are in fact the object of this research. In the next section, I discuss a number of representative excerpts from the audiorecorded interviews, supplementing this information with interview prompts from the MZEE forum as well as additional forum posts corroborating the language attitudes (and underlying ideologies) expressed in the interviews.

7.6 Language attitudes in the German hip hop community

In this section, I present the qualitative analysis of conversations and statements made both on- and offline, centering on a number of different topics. The common thread here is that all of these data excerpts include one or more *expressed language attitudes*. As discussed in the introduction to this chapter, these attitudes do not individually indicate the presence of a systematic and enduring language ideology; for this reason, I contextualize each language attitude by reference to either an established language ideology or else by the presentation of additional attitude expressions that suggest a language-ideological origin.

7.6.1 Orientations toward German and English

A number of interview prompts produced responses including expressed attitudes about German, English, and the use of anglicisms in German more generally—these attitudes, which consisted primarily of assessments, of, e.g., English as an 'easier' language than German, have much in common with the sort of attitudinal dimensions that researchers like Preston (e.g., 2003) engage with. As such, these attitudes deal primarily with perceived inherent properties of the German and English languages.
In the first excerpt, Kranker Karl, a 30-year old underground rapper in Hamburg, has just mentioned that he began (as a teenager) by writing his raps in English, but shortly thereafter switched to German, for which he here provides justification.\(^66\)


**KK**: The thing is just simply, umm, I would say at this point, that I find that it- I in Eng- so that English is simply a more suitable language? Because, so, just from the entire soundscape, uh, that it's, uh, just essentially, um, yes, essentially more harmonious, than, uh, German, already just through, through these, all these, all these sharp sounds in the German language it's already partially, you notice s- one notices that, particularly when rapping faster, that it's just much easier to [unint.] in English. But? Umm, I find. [pause] If you want to really further yourself, as an artist? You have to take hold of the mother tongue, you can't do otherwise no matter how good you control a language, it's just not the same. This feeling, that one- [pause] this feeling that one wants to communicate, I find that it can only be communicated in the mother tongue and in Germany it also just doesn't work [to rap in another language/English].

Karl expresses several attitudes here toward English and German. English is described as *harmonisch*, 'harmonious' and German sounds as *scharf*, 'sharp', an attitude that will be echoed by several other interviewees. English is also seen by Karl as a better language for rapping faster—and the statement seems to be, in essence, a sort of excuse for the use of German. Karl concludes that rapping in a non-native language is not the same, and German, after all, is the *Muttersprache* or mother tongue. The suggestion that English is effectively more suitable for rapping is unexpected, especially in light of the fact that (from my perspective) Karl's rapping style (compared to other German rappers) is rather harsh and aggressive; if German has a sharper, less harmonious sound, it seems that this would provide an advantage.

---

\(^{66}\) 'Kranker Karl', like the remainder of the names given here, is a pseudonym. Some of these pseudonyms were proposed by interviewees themselves; I only created pseudonyms for those who declined to provide their own.
While not all interviewees were rappers like Karl, assessments of the perceived inherent qualities of German and English were nonetheless frequent and were elicited especially well by asking whether the interviewee preferred to listen to American or German hip hop, and why one might be preferred over the other. Especially striking in (2) and (3) is the repetition of similar perceived attributes for English and German, even though in (2) Klaus, the 19-year-old producing student from the HHA, prefers American hip hop, and in (3) Janet, the 27-year-old hip hop fan, prefers German hip hop.

(2) M: So, ehm, wieso hörst du, so, mehr, amerikanisches Hip Hop? Worauf liegt es?

M: So, uh, why do you listen to, like, more, American hip hop? What's the reason?
K: Um, I like the language much better? I find the language to be much more flowing, more flowing, not so choppy like German, German is just really [pause] choppy, very, yeah, angular, you know? [unint.] English is just, an easier language, a more flowing language, and, it simply sounds nicer.

(3) M: Warum, dann, hast du so, deutsche Hip hop, also, gern-lieber als amerikanisches?

M: Why, then, do you like, like German hip hop, so- better than American?
J: Um, I just find [pause] I find German hip hop somehow so [pause] it's just something t-I- I like German hip hop, I think, because umm [pause] because I think it's cool, how the- I understand the wordplay better than in the American, but with the American it's just, I mean, it's just from there, and, you can simply say things a lot more easily somehow in English than in German. And I find the flow simply better, so, it's just, the all in- they just have it m- they just have more to offer.

In (2), Klaus clearly rates English as *flüssig*, 'flowing', *einfach*, 'easy/simple', and as a language that sounds nice (*klingt schön*). German, by contrast, is characterized as *abgehackt*, 'choppy' and *kantig*, 'angular'. While Janet begins by explaining why she prefers German hip hop to American hip hop, it is a utilitarian argument for understanding the wordplay better as a German native speaker—and crucially, not some inherent positive quality of German. The response turns around
to focus on American hip hop, and some perceived attributes of English similar to those in (1) and (2) are brought forward, revealing an underlying positive attitude toward English; it's easier (leichter) to say many things in English, and Janet uses the anglicism Flow to highlight another contrast between the languages (with English still the clear winner).

These sorts of attitudes toward English and German are not only found in the interview data. In a thread on the use of anglicisms in hip hop from the MZEE forum, the following post appears:67

(4) [Post 149635, User 000668, 16-06-2005, 16:16]
Meine subjektive Meinung:
I) es gibt Wörter, deren Bedeutung nicht 1-1 übersetzt werden kann: Styles kicken, bitches, pimpeln, husteln, chillen....
II) allein schon vom Klangbild her ist englisch eine rundere flowigere (-noch son wort wie unter I. beschrieben) Sprache als Deutsch (stark betonte harte konsonanten etc.)
Von daher finde ich nichts schlechtes an einem bewussten Umgang mit der Englischen sprache, sei es auf ganze lieder angewandt (dann aber bitte akzentfrei!) oder in Form von Anglizismen.
Just my 2 cents ;)

My subjective opinion:
I) there are words, which have meanings that can't be translated 1-1: Styles kicken, bitches, pimpeln, husteln, chillen....
II) Even just from the sounds, English is a rounder, more 'flow'ing (-yet another word like those described under I.) language than German (strongly emphasized hard consonants etc.)
For this reason I don't find anything wrong with deliberate contact with the English language, whether it's applied to entire songs (but then please without an accent!) or in the form of anglicisms.
Just my 2 cents ;)

Here we have multiple statements of language attitude to consider. First, there is the implied invocation of the 'luxury loan', the idea that some English words have exact or suitable translations in German (which should then be used; the use of the English word in this case being the 'luxury'), but other words do not, and thus constitute legitimate or necessary borrowings. This is also a recasting of the view that borrowing is motivated by either need or prestige (Hock & Joseph 1996). Interesting here is that the anglicism flowigere (roughly 'more flow-y'), which the forum user points out as an untranslatable borrowing, might easily be replaced by the native form

67 While anglicisms appear in these excerpts, they are not reproduced in bold typeface because of the likelihood first, that this would first predispose the reader to undue focus on specific lexical items, rather than the attitudes expressed in the text (salient anglicisms are discussed when warranted), and second, that bolding might suggest particular stress or emphasis on words where no special emphasis exists.
flüssigere 'more flowing', which, in its non-comparative form, is in fact used to describe English in (2). Also important to note is that this use of a declined adjectival form of English flow expresses an attitude toward language in an iconic fashion, reinforcing the overt commentary on language. The use of English is also perceived as rundere 'rounder' which should be contrasted with the description of German in (1) as kantig 'angular'. There is also a similar characterization of German sounds; harte 'hard' consonants in a statement here reminiscent of scharfe 'sharp' sounds in (1).

In (1-4), then, we begin to see the kind of corroboration of attitudes about English and German that suggests an ideological basis; English is described as aesthetically pleasing, flowing, round, and easy. German, on the other hand, is sharp, hard, angular, choppy, and is described in general as a less aesthetically pleasing language; in general, then, (1-4) seem to take a positive attitude toward (at least the aesthetics of) English and anglicisms and a rather negative attitude toward German, in spite of the differences in communicative goals.

This negative attitude toward German aesthetics, however, does not extend to all other aspects of the language. Janet, 27, also describes German as a beautiful (schön) and useful language. Nevertheless, in the same statement, Janet expresses the attitude that anglicisms are super, and gives a more nuanced explanation of why some words (in this case, chillen, which is also mentioned in (4)) are simply untranslatable:

(5) M: Also, eh, was hältest du von die Verwendung Anglizismen in Hip Hop Kultur hier im Deutschland, oder Englisch, im großen und ganzen?
J: Also, ich, ähm, ich studiere halt auch Deutsch? Deutsche Sprache, so? Und Englisch. Und, ich finde, ahm, [pause] ich finde die Deutsche Sprache total, es ist eine schöne Sprache, mit der du auch [unint.] sehr viel machen kannst, aber, für den Alltagsgebrauch, finde ich das mit den Anglizismen, super. Also weil, es ist genau wie ich sage, ich finde, du kannst sag- manche, manche Sachen sowie 'chillen' zum Beispiel, es gibt einfach kein deutsches Wort, dass das so, du würdest dann 'entspannen' sagen, aber es würde immer noch irgendwie was fehlen, also so 'geil abhängen', mit- also, 'chillen' impliziert so viel, einfach, und die Leute wissen genau was gemeint ist, und das gibt's meine ich, mit vielen Wörtern.
M: So, uh, what do you think of the use of anglicisms in hip hop culture here in Germany, or English, in general?

J: So, I, um, I also study German? German language, right? And English. And, I find, um [pause] I find the German language totally, it's a beautiful language, with which you can also [unint.] do a lot of things, but for everyday use, I find that with the anglicisms outstanding. So because, it's just like I'm saying, I find, you can say- some, some things like 'chillen' for example, there's just no German word, that does- you would then say ' entspannen' [to relax], but something would still be missing somehow. So like, 'geil abhängen' [to hang out in a cool way], with- 'chillen' just implies so [unint.] much, and the people know exactly what's meant, and that's the way it is, I believe, with many words.

Janet’s comment suggests another reason for using English, in addition to aesthetics: some connotative component is lost in the translation ('something would still be missing somehow' / '[it] just implies so much.') This, then, can be seen as a corroboration of the earlier idea in (4) that some words, even when not describing technological innovations, new flora & fauna, etc., can fill a sort of 'gap' in the language or, at the least, express something that cannot felicitously be expressed in German and are thus legitimate/necessary borrowings. This observation is brought into high relief in an interview with Michelle, a 23-year-old hip hop fan and university student. In (6), I ask about the use of certain hip hop-related anglicisms and candidate German translations:

(6)  
M: Oder 'beef'?
Mi: 'Beef' sag ich ja- sage ich ja schon immer.
M: Was ist dann, was ist dann die, große Unterschied in so, ahm, Bedeutung zwischen 'Beef' und 'Streit'?
[long pause]
Mi: Es gibt keinen, es gibt keinen.
M: Und 'Battle' und Kä- 'Kampf', so?
Mi: Ja, da gibt's schon einen, find ich.
M: Ja?
Mi: Ja, wei- weil der 'Battle', das ist ja, meistens- ist ja mit Spaß verbunden,
M: Oder 'Wettbewerb', dann, 'Battle' und 'Wettbewerb', dann?
[pause]
Mi: Ja, gibt eigentlich doch keinen-[laughter]
M: Or 'beef'?
Mi: I say 'beef' all- I say it all the time.
M: What's then, what's the big difference, then, in like, uh, meaning between 'beef' and 'Streit'?
[long pause]
Mi: There isn't one, there isn't one.
M: And 'battle' and 'Kämpf', like?
Mi: Yeah, there's one, I find.
M: Yeah?
Mi: Yeah, be- because the 'battle', that's, well, mostly- is connected to fun.
M: Or 'Wettbewerb', then, 'battle' and 'Wettbewerb', then?
[pause]
Mi: Yeah there's not really one-
[laughter]
Mi: True. [unint.] True, [unint.] because it just sounds cooler, also. Simply just- simply just also for fun. And because one's used to it.

For the word pair Battle and Wettbewerb 'contest', Michelle explicitly states that there is no difference in meaning, but suggests that the English terms are used essentially because they have a connection to fun (Spaß) and they sound 'cooler'. Finally, the words are also established—this echoes Janet's assertion that with chillen, 'you know exactly what's meant'. These attitude expressions can be compared to Piller's (2001) findings that English indexes a 'fun orientation' in German advertising. (5) and (6) above also reinforce the notion that prestige (Hock and Joseph 1996) plays a role in borrowing, and that this role is in fact overtly noted. Coolness, after all, could well be considered a social meaning—compare Eckert's (2008: 465) treatment of 'toughness'—conveyed by the use of certain stylistic resources that carry local (in the social-network sense, if not in the geographic sense) prestige.

Excerpts (1-6), then, highlight a number of language attitudes, several of which I suggest here are ideological in nature, due to the similarity of their reproduction by multiple actors. First, there appears to be a consistent ideological difference in the characterization of German and English; German is found to be aesthetically displeasing as highlighted by the operation of metaphors conceptualizing language-as-substance—German is seen as rough, sharp, angular, and choppier or less-flowing (only Janet's suggestion that German is 'beautiful' stands against the rest of the conceptualizations). English on the other hand, is not only seen as aesthetically pleasing (round, flowing) but additionally as an 'easier' language especially with regard to its use in hip
hop—although it is not particularly clear at this juncture whether this is in reference to the perceived ease of learning, or perceived ease of use—both, in fact, seem likely, in light of the data presented in the next section. The second language ideology evident in (1-6) is reflected in statements that suggest the necessity of borrowing, e.g., chillen, due to the inability to express the same concept in German—some borrowings are, then, conceptualized as legitimate or authorized due to their filling of a conceptual gap in the German language. I will refer to this as the ideology of the 'necessary borrowing'; it interacts with the contrasting and established concept of the 'luxury loan', both of which appear again in the next two subsections.

7.6.2 Anglicism use, youth, and the notion of a 'hip hop language'

To this point, I have discussed attitudes and underlying ideologies about the German and English languages, along with anglicisms in German. To what extent, though, does there exist the perception that an anglicism-laden variety of German is perhaps an entirely different language? Erik Efficiency, a Hamburg-based DJ and instructor at the Hip Hop Academy, as well as a 20+ year veteran of the German hip hop scene, provides an entry into this discussion:

(7)  M: Gibt es, deine Meinung nach, eine gewisse 'Hip Hop Sprache', hier in Deutschland?  
EE: Naja, ja, natürlich. Aber das ist einfach, das ist einfach Jugendsprache, nö? Also das ist zum Beispiel schon auch, da komm ich wieder auf meine Zugfahrerzeiten zurück, das ist halt einfach wenn Leute, wenn Leute so zwanzig Jahre älter sind, als man selber, so? die dann, ähh, im Zug sitzen, und irgendwann fragen wo man herkommt, weil sie das Gefühl haben, man redet halt in äh, in irgendeiner anderen Sprache, aber definitiv nicht ihrer. Weil man einfach soviel Worte hat. So, ähm, aber das ist schon einfach regional so unterschiedlich, was total in den Staaten ja sicher auch so is. […] Wie man auch hier, quasi ganz normale [pause] wie heißt das dann nochmal [pause] Dialekt hat, so? gibt es natürlich in der Jugendsprache auch. […] Es ist irgendwann alles so gleich geworden.

M: Is there, in your opinion, a particular 'hip hop language' here in Germany?  
EE: Well, yes, naturally. But it's simply, it's simply youth language, no? So that's for example like, and I'm going back to my train-riding days here, where like, people, when people that are like twenty years older than you are, right? They then, uhh, sit in the train, and at some point ask you where you're from, because they have the feeling, you're speaking in uh, in some other language, but definitely not theirs. Because one has so many words. So, umm, but it's already simply so different, which I'm sure is totally similar in the States. […] like one also here, almost completely normally [pause] what's it called again [pause] has dialects, like that? There's that in the youth language too. […] but at some point it all became so similar.
Erik discusses several points here; first, he equates the concepts of 'hip hop language' and 'youth language', a parallel which will emerge again in later excerpts. Second, Erik reports on others' reactions to youth/hip hop language, and particularly its conception as another language (but definitely not 'theirs', i.e., Hochdeutsch). Third, Erik comments on his perception of regional variation in youth/hip hop language in Germany.

I continue the analysis of attitudes toward anglicisms by discussing three posts from an off topic anglicism thread on MZEE.com, following each with representative reactions elicited in spoken interviews.68 The posts were read to interviewees, who were then asked to comment on the statements therein. Interviewees sometimes chose to read the posts directly from the interview protocol as well. In the following, I will discuss both the attitudes expressed in the forum posts and the attitudes expressed in interviewees' reactions.

(8) [Post #1 from off topic thread on anglicisms, 09-15-2003]
Ich krieg die Krise.....Ich seh MTV und dann sagen die: "Lasst die Crowd ausflippen mit eurem Sound, gewinnt 2 Tickets und rockt backstag." […] Keine deutschen Wörter mehr sondern nur noch Anglizismen, Anglizismen, Anglizismen….Da isses kein Wunder, dass die deutschen Schüler zu blöd zum Scheißen sind....

I'm losing it….I watch MTV and then they say “Let the crowd flip out with your sound, win 2 tickets and rock backstage.” […] No German words any more, instead only anglicisms, anglicisms, anglicisms….That's why it's no wonder that German students are too dumb to shit....

This post, which reflects a fairly regularly repeated view of anglicisms, overtly connects anglicisms to the purported low intelligence of German (grade school) students. It is important to note that anglicisms are associated here not with the 'elite, modern' conception of English (Piller 2001) but are, in particular, connected with youth culture. In ethnographic interviews, one of the most interesting reactions was that of Bo, an 18-year old DJ and breakdancer at the Hip Hop Academy.

---

68 Because these posts are not part of the MZEE.com corpus, which is drawn from another subforum, usernames are not provided and the posts are numbered 1-3. My interview protocol contained four posts in all, but only three are discussed here.

Bo's statement begins by expressing skepticism that anglicisms have much effect on education, instead placing anglicisms and purported speech problems as part of a larger constellation of problems related to socioeconomic status (‘problem neighborhoods’), social class (‘higher people’), ethnicity (‘don’t have German parents’), and hip hop (‘they're into hip hop’). This is not an isolated concern; another MZEE thread is titled Sprachprobleme der Hip hop Jugend, 'language problems of the hip hop youth.' The key issue is that while Bo masters the ability to switch between an anglicism-containing hip hop variety and a more standard variety, a number of youth can’t—and they are less employable because of it.

69 This thread was accessible at [http://www.mzee.com/forum/archive/index.php/t-100080845.html] as of 11 October 2011; it has since disappeared from the site (older threads tend to go missing with regularity) and is unfortunately not part of the 'Hip Hop Diskussion' subforum and thus not saved as part of the MZEE corpus.
Georg, 26, a Heidelberg-based host of an all-German hip hop radio show, shares these concerns:

(10)  


G: [...] But with anglicisms [pause] taken separately I think, it's OK, because there are still people who speak the clear German that is unadulterated, yeah, in the hip- [pause] Yes. Sometimes I find it negative, with hip hop listeners, especially with youth, that are like 15, 16, that only listen to English, their speech is already very strongly influenced. And then, the natural question is, to what extent do they command the standard language, naturally. Do they know it? So they- if they already know the German language, I think, they know the German words and they naturally aren't using them. And s- then not only, when they're talking about hip hop, but in general language use. Because then of course, their clique? Their friends that listen to the same music, that speak the same language, and then you're naturally not speaking only musically with these concepts [...]

Georg's statement contains a number of language attitudes: First, anglicisms are not seen (in and of themselves) as a major concern in mainstream society, because speakers of 'unadulterated German'—a clear link to the standard language ideologies discussed in the beginning of this chapter—still exist. Georg then goes on to note that anglicisms in hip hop culture are, in fact, a personal concern of his. He connects anglicisms and a hip hop influenced speech to specifically teenaged youth, and raises concerns about their competence in standard German, echoing Bo's concerns in (9). In addition, the concept of luxury loans appears again—'they know the German words and they naturally aren't using them.' Finally, Georg discusses the possibility that hip hop-related anglicisms are not confined to hip hop usage, but instead infiltrate general speech situations, much like Bo's concern in (9) that hip hop youth are effectively unable to style-shift to a more overtly prestigious dialect.

Peter, 22, a beatboxing student at the HHA, repeats the connection of anglicisms to youth language, and in particular teenagers:
Peter connects this use of anglicisms to teenagers ('15, 14-year-old kids') who find them 'cool', but in the same move distances himself (as an older hip hopper, one gets the impression) from this usage and, while he admits to using the word, he uses a German turn of phrase *mit einem lachenden Auge*, a short version of *mit einem lachenden Auge und einem weinenden Auge*, 'with one laughing eye and one crying eye'—an idiom indicating mixed feelings about an action.

The theme that emerges from (7-11) suggests an ideological positioning of anglicisms as connected to (a negatively-assessed) youth language and hip hop culture, although interviewees' attitudes vary regarding the agency of anglicisms in the suggested decline of the German language as a whole. Rather, the (over)use of anglicisms is perceived to be a symptom of larger social problems affecting youth education and jobs. Georg's response, in (10), also indicates an ideological connection to language purism and standard language ideology, and this is perhaps unsurprising given a bit more context: Georg, a radio DJ and graduate student, is unique among interviewees in reporting that he listens almost exclusively to German-language hip hop, whereas most German hip hop fans tend to listen either exclusively to American hip hop or (more commonly) to both German and American hip hop music.

7.6.3 Anglicisms and the fate of German

In reviewing the collected data, one concept in particular reverberated throughout the interviews and forum posts—the notion that a continuing influx of anglicisms might lead to the decline or
even the disappearance of the German language after a certain amount of time (in this case, 100 years):

(12) [Post #4 from off topic anglicisms thread, 09-15-2003]

Ganz genau! Ich find wirklich unglaublich! Wirklich überall, "Let's go, zu großen Family Party mit vielen Boys und Girls in deinem Alter, finde neue Friends"

ich hab mir schon mal ernsthaft darüber Gedanken gemacht, ob in 100 Jahren noch irgendjemand in Deutschland deutsch spricht... Obwohl wir ja laut einer These schon in 50 Jahren mehr Ausländer, als Deutsche in Deutschland haben, aber das is was anderes... Warum machen wirs nicht so wie die Franzosen? Die haben das staatlich geregelt... schade, wenigstens die schlimmsten Aglizismen [sic] könnte man doch wohl öffentlich unterbinden. Wenigstens Anglizismen, für die es auch deutsche Wörter gibt, sollten "verboten" werden...

Exactly! I find it completely unbelievable! Really everywhere, “Let’s go, to big family party with many boys and girls your age, find new friends”

I've already thought seriously about whether there will still be anyone in Germany who will speak German in 100 years... Although according to one suggestion, in 50 years we'll have more foreigners than Germans in Germany, but that's something else...

Why don't we do it like the French? They have it governmentally regulated... it's a shame, at least the worst aglizisms [sic] could then be eliminated. At least anglicisms, for which there exist German words should be 'prohibited'...

This diatribe against excessive anglicism use certainly suggests that anglicisms threaten the continued existence of the German language, and reproduces a strong version of language purism, suggesting specifically that not only do anglicisms pose a threat to German, but that this threat should be responded to by the application of language policy; this can be compared with similar suggestions from the Verein Deutsche Sprache (Spitzmüller 2007). It should be noted that even the suggestion that anglicisms could somehow be eliminated in this way is an ideological statement, and this forum post also ends by reproducing the categorization of 'necessary' vs. 'unnecessary' borrowings, also seen in example (10) above, rating the latter as much worse/more harmful than the former. Finally, an implicit instantiation of Herderian ideology is seen, as the connection between Germany and the German language (and, in passing, even German ethnicity) is evidently being assumed by the poster.

Reactions to this post in the interview data were varied, but none seemed to agree with the statement that German should be regulated, so it is likely that while this view connects to mainstream (or extreme versions of mainstream) language ideologies, it is an ideology
reproduced by only a minority in the hip hop community. Bo's response to this post is seen in (13):

(13)  
B: Nö, denk ich, ne? Ich bin auch eigentlich ganz OK mit Anglizismen, ah, [unint.] Stimmt, Franzosen, und, ah, Italiener, und Spanier sprechen ganz wenig Englisch, und fast auch gar keine [unint.], die sind also stolz auf ihre Sprachen? Ich hab, ahh, ich hab, bin nicht stolz auf Deutsch, so zu sagen? An sich, weil, ich glaub auch die wenigsten Leute stolz auf Deutschland sind, wegen den ganzen Kriegsgeschichten? Also, hast du [unint.] mitbekommen, das wir kaum 'national Pride' haben, so mäßig?

B: Nah, I don't think so, yeah? I'm actually completely OK with anglicisms, ah, [unint.] It's true, the French, and, uh, Italians, and Spanish speak very little English and almost no [unint.] they're too proud of their languages? I have, uh, I have am not proud of German, so to speak? Just for itself, because, I also think the fewest people are proud of Germany, because of the entire war stories? So, haven't you [unint.] noticed, that we don't have much 'national pride', something like that?

In this excerpt, Herderian ideology is still evident. The 'French', 'Italians', and 'Spanish' are invoked as nationalities, and the attributive ihre Sprachen, 'their languages' is presented as unproblematic. Here, it is key to note that while Bo disagrees with the solution proposed in the post, he crucially does not disagree with the notion that anglicisms threaten German, or the suggestion that German will not be around in 100 years. Instead, Bo notes that he's not especially 'proud' of German. Connecting linguistic pride to national pride, he uses the English national Pride rather than the German form Nationalstolz, underlining the extent to which the concept is still somewhat taboo in modern Germany, precisely because of the Kriegsgeschichten, 'war stories'. In doing so, he draws a connection between the language purism evident in the post and nationalist attitudes, distancing himself from the whole. Another HHA student, Keeya, a 19-year old 'nu-style' hip hop dancer, expresses a similar lack of concern:

(14)  
M: Und, ahh, hast du irgendwie so, ein bisschen Angst für die Deutsche Sprache, dass es, so, ehm, das Deutsche Sprache, so, nicht nur weiterentwickelt wird, sondern auch vielleicht zerstört wird durch Anglizismen, oder-

[…]  
K: Ja, i- ja, ich verstehe was du meinst. Das es ein bisschen [unint.], dass es sich so ein bisschen zurückentwickeln könnte, wenn man die deutsche Sprache so an sich sieht. Das könnte natürlich passieren, meiner Meinung nach, ich würde es nicht schlimm finden, aber es, also, wenn ich mir überlege, dass es vielleicht in den nächsten hundert Jahren nochmal so weiter geht? Dann kann es schon sein, dass, [pause] dass sie vielleicht ein bisschen verloren geht […]

161
M: And, umm, do you have, somehow, a little bit of fear for the German language, that it, like, umm, the German language, like, will not only [not] develop further, but also might be destroyed through anglicisms, or, [...] 
K: Yeah, I- Yeah, I understand what you mean. That is a little [unint.], that it could regress a little bit, when you look at the German language in and of itself like that. That could naturally happen, in my opinion, I wouldn't find it bad, but it, so, when I consider, that maybe in the next hundred years goes further? Then it really could be the case, that, [pause] that it maybe is a little lost [...] 

Keeya goes on to give a few examples that come to mind, including a reduction in the number of TV show and movie titles that are translated into German (being carried over wholesale or adapted from English instead), and continues:


K: For us- for the younger generation, I don't find it bad? Because we young people, we deal with- a lot w- with English somehow, because right now it's become a world language, for my grandma, for example, it's very hard. [...] She naturally is more likely to have to think about it, because that's still a completely different generation. I think, it somehow has to go with the times. 

In (14), Keeya expresses agreement with the idea that German could, in some way, be damaged by the influx of anglicisms, but with the reservation that this is only the case if one sees 'German for itself like that', i.e., considers it as a monolithic structure. While Keeya wouldn't really assess the suggested decline of German negatively, she does note that if an influx of anglicisms continues for 100 years, German could be lost. Finally, Keeya expresses concern not for German, but for the elderly and their ability to comprehend a changing language. This excerpt, then, shows little evidence of standard language ideology (German is only reluctantly considered an object in and of itself) or Herderian ideology. English is conceptualized as a 'world language' and again associated with youth. Keeya is also a member of a visible minority in Germany, and the marked difference between her attitude and other respondents' attitudes may not be unrelated to an experiential difference in identification with German. However, making generalizations about the possible correlation of ethnicity with language attitude or ideology is not warranted in this
case as this was not the focus of the interviews, and also because of the small number of interviewees.

On the question of anglicisms and German language decline, Georg, the radio DJ, speaks directly:

(16) **M:** [reading voice] *Deiner Meinung nach, also, bedrohen Anglizismen die deutsche Sprache? Die deutsche Identität?*

    [long pause]

**M:** [reading voice] In your opinion, then, do anglicisms threaten the German language? The German identity?

    [long pause]
    **G:** I think that the German language loses through the fact that there are many anglicisms, yes, and? [long pause] Yes, and in large parts there's no consciousness of the fact that one is using a lot of English [words/phrases], and loses German counterparts. I don't find it problematic in modern language, and internationally there's in th- In the computer field, or informatics I don't find it problematic. I find it problematic when it's found in everyday speech, when too m- too much is adopted, that [unint.] I find the problem, where I think that the German people a truly large consciousness have for, right [unint.] that there are many cultures, for example Turkish, then you also have other influences on the German language, that means, not just this one-sided, that it's just coming from English, but rather a general problem that the German language is being improved, and partially through dialects, loses its [purity/clarity], partially through Turkish influences, partially through anglicisms.

Here, Georg expresses serious concern about the future of the German language, and views anglicisms as only one source of influence (the others being Turkish and German dialects). The specific concern that German dialects are playing a role in the loss of *Reinheit* underlines that this concern is not for the German language including its dialects, but rather, for the loss of an idealized, standard German, which is furthermore described in terms of its *Reinheit.*

163
here raises a difficulty in translation and interpretation—it can mean either 'clarity', a concept less directly connected to language purism, but it can also indicate 'purity', in which case a direct connection to language-purist ideologies is evident. In any case, Georg expresses attitudes indicative of both a standard-language ideology and the notion of subtractive or disruptive borrowing, indicating that the use of English is directly connected to the loss of German words. As a final note on the translation, I translate the phrase dass die deutsche Sprache verbessert wird, as 'that the German language is being improved', which seems out of place in the overall discourse (given that it is explicitly stated as a 'problem'). The reason is that verbessern in this context is likely used in a somewhat ironic manner, and 'forcibly improved' might be a more accurate translation of the connotation here.

Another post from the off topic forum seems to take a position more like the statements of Bo and Keeya:

(17) [Post #2 from off topic anglicisms thread]

Sprache lebt nun mal, sie entwickelt und verändert sich. Wenn irgendwelche Betonköpfe kommen, die auf Teufel komm raus die deutsche Sprache retten müssen, dann kotzt mich das an.
Freut euch doch. Je mehr andere Sprachen einen Einfluss auf die eigene haben, umso weniger fällt es schwer, fremde Sprachen zu lernen. Ich habe nichts dagegen, wenn die Menschen auf der Welt in hundert Jahren wieder eine gemeinsame Sprache sprechen.

Language lives, it develops and changes. If some hardliners come along that need to save the German language come hell or high water, then that makes me sick. Be glad. The more influence other languages have on your own, the easier it is to learn foreign languages. I don't have anything against it if, in a hundred years, all people on the world have a common language again.

This forum post begins by espousing a belief that most socio- and historical linguists (who are, as Spitzmüller (2007) notes, ideology brokers) would heartily agree with—and one which is silenced in expressions of mainstream/standard language ideologies: the notion that language change is inevitable (and not negative). The poster, however, then treads into linguistically dubious water, expressing the possibility that languages could essentially converge into a 'world language' within a century. What is not stated outright here is that this outcome would entail the loss of the German language as distinct, but as seen in (12) and (16), the positive assessment of such an outcome is contested.
For some in the hip hop community, then, a rather conservative view of the German language echoes the set of concerns tied up in standard language ideologies which operate in the German sphere. For others, however, the use of anglicisms is not considered problematic—but this is only because these speakers won't mourn the 'inevitable' loss of German. The common theme from (12-17) is that none of the expressed attitudes consider a possible future in which the German language incorporates anglicisms while still remaining, in essence, the German language. Crucially, both sets of attitudes expressed here—that anglicisms are a problem because they threaten the German language and that anglicisms are not a problem because the loss of German isn't something to mourn—rely on the tacit assumption (a shared and enduring belief and therefore part of an ideology) that the German language is necessarily in decline at least in part because of English influence.

7.6.4 Anglicisms and novelty

In (11) above, Peter reacts to an excerpt containing a number of anglicisms with scorn, noting that the anglicisms are, in effect, 'speaking to these 15, 14-year-old kids, that then find it cool'. This statement is of particular interest in light of the findings in Chapter 5, specifically that frequently-used anglicisms are likely to experience a drop in frequency over time. A final pair of excerpted statements about anglicism use from Peter and Erik directly illuminate some motivations for that finding:


P: I have already noticed, that a lot of people pick out for [unint.] themselves, like, a few [English] words? But I think, that it's not, so it's not, I think, so extreme. So, it's maybe already- it has happened, that one then says some English word or another somehow for like- for like half a year, but that's it, in any case.

In (18), Peter directly addresses the way in which novel anglicisms are used, and there are two components of his statement that I would like to call attention to. First, sich raussuchen, 'to pick out for oneself' suggests that novel anglicisms are first used as stylistic idiosyncrasies by individual speakers (recall here some of the 'personalized' spellings and capitalizations of peace
in Chapter 5). Second, the short-lived nature of anglicism usage is highlighted by the statement that 'one then says some English word [...] for like half a year, but that's it'. Erik's statement in (19) further illuminates the process of anglicism adoption:


EE: For example. It's the case that I- I for example, for a certain time always said, like, 'Ja, rocket!', like that? Just for the hell of it! So, when I could also say, 'Klar, alter!'. But I just didn't have any interest in 'alter' anymore.

Erik goes on to explain that he picked the phrase up from the German-dubbed version of the 1988 American movie 'Colors', in a scene where two gang members are preparing to perform a drive-by shooting. Particularly interesting here is the suggestion that the novelty of this phrase—rather than the idea that it could express something inexpressible in German (Erik mentions Klar, alter! 'yeah, man!' as fulfilling an identical pragmatic function)—was the key motivation for its use. Second, Erik mentions that he lost interest in the slang address term alter (which is similar in its use to digga, see fn. 62). Alter is a native German term, and not an anglicism, but its use in this context is, essentially, slang—it fulfils the requirements laid out by Dumas & Lighter (1978) in that it lowers the dignity of the occasion in which it is used, and expresses a special familiarity with the referent (this can also be construed as expressing solidarity). Erik's loss of interest in alter, then, and move to the idiosyncratic quasi-borrowing (Ja, rocket!) can be ascribed to the drive for novelty in slang; and for the hip hop community, the American hip hop vocabulary provides a large number of novel candidates can be found. However, the process is seen to come full circle when we return to expressed attitudes like those in (11), where it is obvious that some of these forms, especially those perceived to become popular among teenagers, are eventually discarded.

7.5 Conclusions and implications

In the analysis presented in this chapter, each of a set of ideologies which I have called the standard language ideology complex (which includes the ideology of the standard language, language purism, and Herderian ideology) came to the fore in interviewees' attitudes toward German. The analysis here, however, reveals a larger and more complex system of enduring and
shared ideologies which interact with these and with each other, and which operate with regard to the use of anglicisms and in orientations to English within the hip hop community. Yildiz (2004: 322-323) poses an excellent formulation of the themes which I address in this chapter:

Is it then “English” which puts pressure on “German” or “German” which borrows from other languages for its own purposes? I suggest that the most important question we need to ask in this respect is: what desires does the use or evocation of English satisfy for German speakers? What does “Americanization” look like from this vantage point? What functions does it fulfill? The interaction between languages does not take place unilaterally. Therefore, we need to alter our framework to register fully the processes in which the German language is involved. Even if we assume the dominance of English in general and Americanization in particular, we still need to consider local appropriations of English, as well as other languages, and what they tell us about local forms of agency and desire.

The first ideology identified through qualitative analysis of interviews and forum posts is a surprisingly consistent language-aesthetic ideology, which encompasses several routinized dimensions of difference between the perceived innate properties of English and German. Under this ideology, English is seen as a more aesthetically pleasing language than German, which is characterized as aesthetically displeasing. Particular dimensions of aesthetic assessment which are commonly repeated under this ideology are shape, fluidity, and sound (round vs. angular, flowing vs. choppy, harmonic vs. sharp). Related statements which may be indicative of an ideology, but are not well-attested enough in the present analysis to be sure, characterize English on a utilitarian, rather than aesthetic, basis, as 'easier' and 'more suitable for rap'.

The second ideology identified here is stated directly in several of the excerpts, but also found as an underlying assumption in others. This is a dualistic ideology about language contact, which I will call the necessary borrowing/luxury loan ideology. In essence, it concerns the distinction between two types of borrowings: those that are untranslatable into German, which thus legitimately fill some manner of logical gap in the language, i.e., referentially-necessary borrowings which are perceived to convey terms' descriptive content (necessary borrowings, assessed as neutral or positive) and those that do not fill such a role or do not differ in semantic content or pragmatic function from perceived German equivalents; these, then, are the luxury loans, and are in general assessed negatively. In my capacity as a linguist (keeping in mind that I am in some way still an ideology broker), it seems that this distinction is hard to maintain on objective or scientific criteria. It is true that some anglicisms, especially those which have to do
with technical concepts that entered German alongside hip hop culture, do not have commonly-used German equivalents or even calques, and as such, the borrowing is motivated by denotational need. Taking one such term, like breakdancing, as an example, one could readily, if clumsily, calque it as, e.g., bruchtanzen (a Google search returns some 300, largely tongue-in-cheek hits). The failure of a calque in this case, however, is crucially not that it does not convey some part of the referential meaning, which can, after all, be arbitrarily assigned to a number of candidate neologisms or calques, but rather that the English word simply sounds better or more authentic to users—thus, there is additionally an expressive motivation for borrowing. A purported luxury loan like chillen or beef, on the other hand, could essentially be considered to convey only additional connotational and expressive content when compared with the native terms entspannen or Streit, respectively. Janet and Michelle confirm in (5) and (6) that there is something extra going on with the connotations of the English terms. There does not, then, seem to be an inherent logical distinction between the necessary borrowings and the luxury loans. Even in the case of a denotational gap in the language, the decision to borrow rather than to calque or coin a novel German word is seemingly based on some measure of expressive or connotational desire associated with English—and thus, in terms of denotational meaning, borrowing is never a strictly necessary or inevitable outcome, even if there do exist folk categories (of which anglicism membership seems to vary from individual to individual) intuitively understood by language users.

The third ideological complex that is identified in this analysis is instantiated in views that anglicisms are indicative of a non-standard language variety, which is alternatively or simultaneously identified as hip hop language and/or youth language. This non-standard variety is, in general either negatively assessed or perceived to be negatively assessed (by others); this is a result of the interaction of this ideology, or else the reliance of this ideology upon, the standard language ideology. Multiple commentators also draw a link between this anglicism-containing, non-standard hip hop/youth language and lower classes or non-German ethnicities. While the term was not used in interviews, possibly because of its status as a taboo/potentially offensive designation, there exists a great deal of agreement between the characterization of non-standard varieties in the interview/forum data and popular conceptions of a minority-based Kanaksprak (Yildiz 2004). From my perspective, parallels can be drawn between this configuration of non-
standard varieties/ethnolects/youth language in the German sphere and the influence of (or appropriation of) African American English on (in) trans-ethnic youth language and popular slang in the US (for treatment of which see, e.g., Bucholtz 1997).

Finally, the conception of anglicisms as inevitably deleterious to the German language constitutes another distinct ideological component visible in the ethnographic data. This is made evident through a sort of shared story that emerges from different assessments of concern about the future of German. While interviewees and forum users do not, in general, agree on whether the decline or development of German through English influence is a good or a bad thing, there is a shared underlying assumption that a continued influx of anglicisms will eventually cause the decline or destruction of the German language. This ideology interacts with the ideology of language purism in that it equates foreign inclusions with impurities; it additionally informs the ideology of the necessary borrowing and luxury loan, ultimately suggesting a subtractive/disruptive view of language contact.

The ideologies identified above, in general, tend to interact with the set of standard language ideologies; while the assessment of potential outcomes may be contested within the hip hop community, a number of standard-language and language-purist assumptions seem to underlie the attitudes expressed by members of the German hip hop community. Crucially, this runs counter to the hypothesis that these ideologies would generally be resisted as part of the conception of hip hop as an opposition culture. Speaking generally about the excerpts above, the views expressed are surprisingly congruent with the standard language ideologies. While the English language is generally valued as a more aesthetically pleasing language than German among members of the hip hop community, views on the mixture of English with German, i.e., the use of anglicisms, are sometimes surprisingly and strongly negative. To underline the extent of this congruence, once can find definite parallels between almost every sentence of the excerpted views expressed by Transportation Minister Ramsauer in his stern interview (Schütz 2010), discussed in Section 7.2.4, and material in the excerpts from the ethnographic research here, from the Herderian equation of nation and language in the MZEE thread in (12) to Keeya's concern about the disruption of intergenerational communication in (15) to Georg's assertion that Germans do not give enough consideration to their language in (16).
While it is extremely surprising to find reproductions of the standard language ideology complex within the hip hop community, this is only part of the story. In (5) and (6), Janet and Michelle positively assess the use of anglicisms without reservations, and much of the data, like Keeya's statement in (14), indicate a general lack of concern about the fate of German. Within the German hip hop community, it does seem that anglicisms are acknowledged as having both communicative and social utility, but that an increased or abnormally frequent use of anglicisms, especially by lower-class Germans, teenagers, immigrants or by mass media outlets targeting this demographic, is assessed negatively and thought to indicate a lack of education in the former case or inauthenticity and pandering in the latter.

In closing, one set of attitudes in the present study might well be described as entirely subcultural and directly speaks to the finding of a negative correlation between anglicism frequency at an initial time and change in anglicism frequency discussed in Chapter 6. The attitudes discussed in section 7.4.4 underline the notion that hip hop community members prize novelty in language (and discard, on an individual or group level, forms that no longer fulfill the requisite stylistic functions). A more thorough discussion of the relationship between the anglicism dynamics and community attitudes about anglicism novelty, as well as the implications of the ideological diversity in the German hip hop community, is the focus of the next and final chapter of this dissertation, Chapter 8.
8.1 Overview

In this dissertation, I have focused on the research object of anglicisms, or English-derived linguistic material, in the German hip hop community. In terms of its social impact, this research was prompted by current sociopolitical debates about the influence of English on German, and in particular the question of whether such influence could be considered to threaten the future of the German language. The domain of hip hop was chosen in part because it is a leading example of a globalizing culture, originating in an English-speaking population, in which language is central—leading to a situation where novel anglicisms are more frequent within the German hip hop community than outside. Through methodologically diverse studies addressing a number of aspects of this question both qualitatively and quantitatively, I conclude here that while English influence on German proceeds in a number of interesting and previously unrecognized ways, the extent of this influence is overstated by language purists, and anglicisms are not likely to lead to the decline or loss of the German language, as many commentators suppose.

8.2 Considerations of research object and research domain

After discussing the debates which catalyzed the research project presented in this dissertation and identifying anglicisms as the research object in Chapter 1, I turned to the task of precisely defining anglicisms, in particular discussing the language contact phenomena of borrowing, code-switching, and code-mixing with reference to previous theorization on the subject in Chapter 2. In that chapter, I motivated the use of the term anglicisms as inclusive of a number of types of linguistic material which occur in a German-language context and are English-derived or generally perceived as English-derived. Anglicisms are differentiated primarily on account of
methodological concerns into two major categories, *visible* anglicisms and *invisible* anglicisms
primarily based on the metric of *formal salience*, following Onysko (2007) in this regard. The
former category is the one most relevant to the present research, and is comprised of borrowings
from English (*Beef*, 'argument'), including lexical items from multiword borrowings or 'phrasal
anglicisms' (*in your face*), and instances of code-switching (*New tracks are up on […]*); the
debated distinction between borrowings and instances of code-switching or code-mixing was
determined to be largely immaterial for the present research project (with the caveat that such a
distinction may be relevant in other examinations of language contact phenomena). The category
of visible anglicisms additionally includes borrowings which are nativized through the
acquisition of German inflectional or derivational morphology (*gebattlet*, 'battled'; *flowendere*,
'more flowing') as well as hybrid German-English compounds (*drogendealen*, 'to deal drugs').
Finally, Onysko's (2007) *pseudo-anglicisms*, like *Handy*, 'cell phone', where the English origin of
a lexical item is controversial or unknown, but the form is widely believed to be English-derived,
are included under the umbrella of visible anglicisms. *Invisible anglicisms* include calques
(*doppelzeit*, 'double-time') and semantic extensions (*Frieden* 'peace' as a leave-taking). Because
of the reduced salience of invisible anglicisms, as well as the related methodological difficulty in
identifying them, the category of visible anglicisms was made the research focus of this
dissertation.

In Chapter 3, I introduced Internet forums, from which data for this dissertation were collected,
as a type of computer-mediated discourse. I then described the primary dataset, an original
corpus of German-language hip hop discussions from the Internet forums at MZEE.com ('the
MZEE corpus') as well as a secondary corpus of English-language hip hop discussions from the
Internet forums at Projectcovo.com ('the Project Covo corpus'). The MZEE corpus spans 11
years (March 2000 – March 2011) and is 12.5 million words in length, while the Project Covo
corpus spans 8 years (June 2003 – November 2011) and is 19.4 million words in length. I also
detailed in Chapter 3 the automatic collection of these corpora, as well as corpus formatting and
initial processing, which removed extraneous material like URLs and separated quoted or
automatically-cited material from the main text. An initial estimate of anglicism frequency in the
text from a by-hand analysis of a small subset of the MZEE corpus, finding that anglicism
wordforms comprised 6.89% of the word types in the corpus, and 7.35% of the word tokens, was
contrasted with Onysko's (2007) findings for the 2000 volume of German newsmagazine Der Spiegel, in which he found 5.8% of the word types and only 1.11% of tokens in that corpus to be anglicisms. This contrast established that the hip hop forum community uses a slightly wider variety of anglicisms, but those anglicisms comprise a much greater portion of the text than mainstream news discourse—and this initial examination of the MZEE corpus also provided a point of departure for the large-scale computational corpus analysis that followed.

8.3 Research findings and implications

In order to provide an overview of the findings of this research project, I reproduce the specific research questions identified in section 1.5 here:

(1) Can anglicisms be reliably identified by automatic means in natural, highly contextualized data (like the linguistic production in online forums)?

(2) How does the nativization of anglicisms proceed, and what factors condition this process?

(3) What factors cause anglicisms to fail or succeed over time, and what does this tell us about their functions?

(4) What are the underlying and motivating ideologies that affect the use (and possibly the adaptation) of anglicisms?

The automatic identification of anglicisms in the MZEE corpus in Chapter 4, which addresses question (1), resulted in a list of 850 frequent anglicisms which subsequently enabled the investigation of a diversity of unexpected linguistic forms, i.e., the question posed in (2), in Chapter 5, and the larger-scale quantitative analysis of anglicism fate, question (3), in Chapter 6. The issue of ideological conditioning of anglicism use and adaptation, the focus of question (4), was subsequently addressed in Chapter 7, illuminating motivations for the linguistic and distributional facts seen in the other chapters.

I will begin, then, by summarizing the development and outcome of the anglicism classifier described in Chapter 4. As mentioned above, the automatic identification of anglicisms by computational means played a major role in making the subsequent analyses feasible; due to the size (roughly 12.5 million words) and chronological length (over a decade) of the MZEE corpus, the by-hand identification of anglicisms in anything but a small portion of the text would not
have been feasible. The development of an anglicism classifier was not only useful in this regard, however: in addition, the development of extensions to the basic $n$-gram classifier in order to capture a greater proportion of anglicisms in the text was invaluable in identifying some of the properties of nativized anglicisms (namely, affixation and compounding) which confounded the classification problem, setting the stage for the later discussion in Chapter 5 of non-standard instances of nativization.

The ramifications of the classifier development detailed in Chapter 4 additionally extend beyond its utility in the context of this study. The success of the classifier in identifying a list of candidate anglicisms of which a mere 8.7% were found, upon hand-correction, to be misclassified German forms demonstrates that the problem of anglicism identification, even in nonstandard texts is, by and large, a computationally tractable one.

The extensions developed alongside the classifier, in particular the compound-cutting and affix-stripping algorithms, while not generalizable to other data in their instantiations here, demonstrated the efficacy of customizing classifiers for nonstandard language identification problems where the data is less standardized or consistent than the traditional data sources like newspaper articles more commonly used in language identification research. These extensions, used with a 1-6-gram classifier, improved performance on the development set—increasing precision on anglicisms (at 95% recall) by 9 percentage points over an unaltered 1-6-gram classifier, and by 50 percentage points over a baseline 1-gram classifier, results which are additionally reported upon in Garley & Hockenmaier (2012).

The classifier described in Chapter 4 is neither the most complex nor the most powerful language identification classifier in existence. However, the customization of the classifier to this sort of data is novel. A final implication of the classifier development in Chapter 4 stems from the fact that it reliably recognizes English material in German natural language production, and this strongly implies that anglicisms in German are, in general, noticeable to speakers of German, whose linguistic faculties are certainly exponentially more complex and powerful than this

---

70 Precision on anglicisms, as defined in Chapter 4, is the proportion of correctly identified anglicisms to all identified anglicisms; recall on anglicisms is the proportion of correctly identified anglicisms to all anglicisms in the text.
relatively simple classifier. An important conclusion that follows from this is that anglicisms do not—with a few exceptions like *dissen*, which fits German linguistic rules at all levels—fly under the radar, so to speak; German speakers are aware of English inclusions in their linguistic input, a fact confirmed by the ease with which interviewees in Chapter 7, and language commentators discussed in Chapters 1 and 7, are able to identify and discuss English borrowings.

In Chapter 5, I selected several examples from the hand-corrected list of anglicisms identified by the classifier with absolute corpus frequency greater than 100 in order to investigate non-straightforward instances of linguistic nativization. Based on wordform variety and distributional facts, I argued that surprising orthographic forms like *batteln* (for *battlen*), *peaze* (for *peace*) and *flowed* (for *flowt*) are not the result of typographical error or the breakdown of German linguistic rules, but rather result from the application of both German and English (morphological, phonological, and orthographic) rules, which can interoperate in various ways. German and English rules can conflict, as in the case of *battlen/batteln*: the former form fulfills the desideratum of faithfulness to English orthographic form, with the addition of final -(e)n to form a German infinitive, but the latter form fits German orthographic conventions and maintains congruence between orthographic and phonological paradigms. In the case of *peaze* and non-participial *flowed*, the explanations are related. While the use of <z> in *peaze* is undoubtedly connected to the grapheme’s stylistic use in youth subcultures, the correspondence between the orthographic form *peaze* and the pronunciation as [piːz] can be maintained if <z> is read with its English value /z/ (rather than its German value /ts/) and if the application of the German productive phonological rule of word-final devoicing subsequently causes the /z/ to be realized as [s]. This solution (the use of <z> in reproducing the word *peace*) carries with it the additional benefit of avoiding the grapheme <c>, which, without a following <h> is uncommon and not well-established in German. The final case, and possibly the most interesting one, involves the apparent use of the English participial suffix -ed with anglicisms, which is straightforward in cases like *sachen die dort released werden*, ‘things that are being released there’, in which the German alternative, *rausgebracht*, would also take participial form. However, the -ed suffix appears in additional morphological conditions, including the syncretic 3rd person singular present (*Der Thread sucked*, ‘this thread sucks’) and the 2nd person plural imperative (*voted für den clip*, ‘vote for this clip’), both of which are regularly expressed in German with the suffix -t.
Word-final devoicing again provides an explanation, as it would allow the \(<d>\) in –\textit{ed} to be realized as \([t]\), as in the German word \textit{Rad}, 'wheel', pronounced \([ra:t]\).

These examples highlight the role which orthography can play in mediating the borrowing process, as well as the utility of orthography for research on the linguistic origins of borrowed forms, the latter point having been exploited for several centuries by philology and historical linguistics but often overlooked in modern sociolinguistic research. In fact, none of these alternations would be apparent in a spoken corpus, but crucially, these examples speak of the detailed interaction between two linguistic systems and how language users understand and apply the rules of those systems.

What the examples from Chapter 5 demonstrate is that borrowed material is not necessarily either straightforwardly reproduced in its original form, adhering only to the rules of the donor language, or immediately fully nativized, adhering only to the rules of the receptor language, but can take forms which involve a negotiation between these two sets of rules—and that this process, even for apparently puzzling surface forms, is not haphazard but proceeds in a systematic fashion. In terms of concerns about the influence of English on German, this argues against conceptions of English influence as invasive or destructive of German rules and norms. Instead, this influence results in hybrid and adapted linguistic forms which, if successful in being adopted, might eventually be unproblematically accepted as part of German. However, the analysis performed in Chapter 6 suggests another reason that these concerns may be overstated: the borrowed forms may not remain for long, speaking on historical timescales.

In Chapter 6, I analyzed the word dynamics of the 850 anglicisms in the MZEE corpus diachronically through two analyses, both of which involve the prediction of frequency change over time on the basis of different predictor variables. In the first analysis, I investigated the relationship between frequency at an initial year-long time window \(T_i\) of selected anglicism stems in the Project Covo corpus, an English-language hip hop discussion corpus, and frequency change for those stems in the MZEE corpus over various lengths of time. I found that frequency in the Project Covo corpus did in fact significantly correlate with later frequency change in the MZEE corpus, and the strength of this correlation (and the coefficient of determination, which
indicates the amount of variance explained by a predictor variable) was highest in a two-year timespan, e.g., change in frequency in MZEE between 2004-2006 was more predictable (from initial frequency in Project Covo in 2004) than MZEE frequency change from 2004-2005 or 2004-2007. The correlations were additionally positive, indicating that anglicism stem frequency in MZEE seems to increase or decrease to some extent with frequency in Covo two years earlier.

While it is not suggested that this is a directly causal relationship, as most MZEE forum members are not assumed to read the English Project Covo site, it does suggest a more general time lag of roughly 2 years between a lexical item's popularity in the US and its popularity as a borrowed form in the German hip hop community, which is perhaps somewhat surprising given the generally near-simultaneous release of hip hop albums in both countries and the widespread use of the Internet to disseminate hip hop-related news and media. However, the adoption of trends, for example, the rise in popularity of a new subgenre of hip hop music, might be expected to lag somewhat—because of the following tendency, articulated by Milroy & Milroy (1985: 362): “Those speakers whose ties are weakest are those who approximate least closely to vernacular norms, and are most exposed to pressures for change originating from outside the network.” I suggest that the transmission of primarily mainstream or the most popular and widespread practices from the US scene to the German scene through globalization would explain this lag in transmission, which is to say that it takes a certain amount of time for new innovations to 'catch on' within each community.

In the second analysis from Chapter 6, three predictors internal to the MZEE corpus, including frequency and measures of word dissemination over users and threads (\(D^U\) and \(D^T\)), were measured at an initial time and compared with frequency change between the initial time window and a later time window (the same response variable as in the first analysis). In this analysis, however, the word dynamics of the full list of 850 anglicisms were compared to those of the remaining (primarily native German) words in the corpus. Results indicated that while the native German words' initial frequencies and values of \(D^U\) and \(D^T\) negatively correlated with later change in frequency, the strength of this correlation was miniscule. However, for the set of anglicisms, stronger negative correlations of frequency and the dissemination measures with frequency change were found—and in the strongest case, initial anglicism frequency predicted
frequency change over a three-window period (time windows in this case were chronologically-ordered corpus chunks of 2 million words each), with initial anglicism frequency explaining 18.9% of the variance in later frequency change. This alone is not enough to suggest that frequency change can be effectively modeled by earlier frequency for loanwords, and a number of additional variables (as well as some means of variable selection) are desiderata for future extensions of this research, as discussed in section 8.3. However, the contrast of this result to the results for native German forms suggests that there is in fact something different going on with the anglicisms, specifically that anglicisms behave very differently in terms of word dynamics than native German words, as increases and declines in the use of anglicisms are much more sensitive to factors like frequency and dissemination. This in turn suggests that these borrowings are in many ways volatile, and the fact that the correlation is negative is also important, as it indicates that frequent anglicisms tend to decline in frequency while infrequent anglicisms tend to rise in frequency.

The implications of these findings are extensive—if popular anglicisms are penalized for their frequency and to a lesser extent for their wide dissemination, this suggests that anglicisms may come into and fall out of use relatively quickly, speaking on historical time scales. While the 'time windows' used in this analysis did not correspond to equal lengths of time, these effects were certainly detected within the space of a decade. Haugen (1950: 211) writes in his seminal treatment of borrowing:

At first blush the term 'borrowing' might seem to be almost as inept for the process we wish to analyze as 'mixture'. The metaphor implied is certainly absurd, since the borrowing takes place without the lender's consent or even awareness, and the borrower is under no obligation to repay the loan. One might as well call it stealing, were it not that the owner is deprived of nothing and feels no urge to recover his goods.

The present findings add a new aspect to this dynamic—while the donor language, of course, loses nothing in the process, many borrowings do seem to be temporary in nature; were Haugen writing today, a more apt metaphor for the process might be found in Internet piracy where a file, say, a song, has been copied from one user to the other (with the first losing nothing in the process). However, the file containing that song may well be discarded by the 'borrower' after its utility and novelty have worn off. This works well as a metaphor for borrowing: while the lexical
items are not removed from the donor language, many of them do not seem to remain in the recipient language for long.

Having identified these tendencies in Chapter 6, I investigated the motivations behind them in Chapter 7, which first related language attitude to language ideology, identifying a standard language ideology complex, comprised of the ideology of the standard language, language purism, and Herderian ideology, in the German-speaking sphere. This ideology complex was identified on the basis of previous research (notably Spitzmüller's (2007) analysis of metalinguistic discourses) and supported by recent media discourses which I collected. I then turned to the investigation of language attitudes indicative of ideologies operant in the German hip hop community, analyzing ethnographic interviews I conducted in Hamburg in the summer of 2010 in tandem with forum posts expressing related attitudes toward German, English, and the use of anglicisms. Through this analysis, I was able to identify each of the components of the standard language ideology complex as circulating within this community (although not held by all members), as well as a number of additional ideologies circulating within the German hip hop community:

(5a) A language-aesthetic ideology, which conceives of languages as substances with innate aesthetic properties: English is described as 'rounder' and 'more flowing', while German is 'hard', 'angular', and 'choppy'.

(5b) An ideology distinguishing between 'necessary borrowings' and 'luxury loans', which predisposes speakers to conceive of some borrowings as filling conceptual gaps in a language, while others (the 'luxury loans') have close counterparts in German. This sense of perceived necessity means that in this distinction, 'luxury loans' are generally negatively assessed.

(5c) An ideology which associates anglicisms with a (negatively-assessed) youth / hip hop language variety.

(5d) An ideology which conceives of anglicisms as inevitably deleterious to the German language (and possibly its continued existence).

(5e) An ideology which assigns anglicisms value in accordance with their novelty; this has some correspondence with popular and academic conceptions of slang.

The last of these identified ideologies provides a direct explanation for the word dynamics of anglicisms found in the analysis in Chapter 6. The assignment of value to novel (and thus
infrequent/not widely distributed) anglicisms explains the negative correlation of frequency and the dissemination measures with frequency change—in a subculture which prizes the novel, anglicisms which are very frequent and well-disseminated would not be expected to maintain their popularity. Conversely, anglicisms which are neither frequent nor widely distributed would tend to function very well as markers of belonging to a prestigious in-group. There do seem to be notable exceptions to this generalization, however, as words like cool have maintained their popularity in German for decades—and the investigation of such well-established anglicisms is a direction I propose for future research in section 8.4 below.

The ideologies stated in (5b-d) are each in some way indicative of linguistic conservatism and linked to parts of of the standard language ideology complex. As discussed in Chapter 7, (5c) connects to the ideology of the standard language, (5b) and (5d) rely on linguistic purism, and Herderian ideology informs the negative assessment of (5b-d), as the perceived decline of German would, under the Herderian equation of nation, people, and language, inevitably spell the decline of the German nation and people.

What was not found in the ethnographic interviews was the hypothesized systematic resistance within the hip hop community to the standard language ideological complex. The ideology stated in (5a), for example, positively assesses English relative to German, but this is in some ways compatible with, e.g., language purism, as it is not anglicisms that are positively assessed, but rather the (unmixed) English language. While a structured, ideological resistance was not found, an interesting point is that the assessments associated with these ideologies did vary within the community. In essence, while there was widespread agreement with the story that the decline of German was inevitable through English influence, hip hop fans differed in their assessment of that hypothetical outcome, and some even voiced their support for the eventual disappearance of the German language. The Herderian ideology equating a single nation, people, and language, then, can in fact be considered weakened, if not outright contested, in the German hip hop community. Returning to (5e) briefly, the valuing of novel anglicisms could also be seen as contradictory to the ideology of the standard language—which prescribes, after all, a homogeneous and idealized version of language.
In terms of implications for the linguistic concerns catalyzing this dissertation, which in effect are stated succinctly as ideology (5d) above, the reproduction of these very concerns in the 'resistance vernacular' of the hip hop community truly displays the depth of penetration of these concerns throughout German society. If, even in the multicultural, globalizing, youth-oriented, US-oriented hip hop culture, the standard language ideology complex is still dominant, and hip hop fans echo conservative *Bundesverkehrsminister* Ramsauer's concerns about language (section 7.2.4), this result should simultaneously assuage the fears of those concerned about the future of the German language and prove an unalloyed victory for the dominant bloc institutions (Lippi-Green 1997: 64) which promote standard language ideologies.

8.4 Directions for future research

While this dissertation represents a definite advance in understanding the process whereby linguistic material is borrowed, the distribution of that material, and the ideologies that condition its use, much ground remains to be covered before a full account of these language contact phenomena can be said to exist. First and perhaps most obviously, it is desirable to investigate not only other language pairs, but also other subcultural communities (especially other youth, Internet, and music-oriented subcultures, so as to be able to state with increased confidence whether some of the findings here might generalize to more or less similar language contact situations, i.e., to what extent the patterns and motivations reported here are in fact unique to language contact in hip hop culture. Second, the interaction of language contact and orthography from a sociolinguistic perspective remains sorely understudied. Future analyses might profitably investigate additional examples of borrowing between language pairs that share a script or from language pairs with very different scripts. The advent of Internet communication and the attendant abundance of written data that can be easily collected would greatly facilitate such analyses.

One of the analyses included here that could be taken much further is the investigation of anglicism frequency change and its predictors. Ideally, a full statistical model with additional variables and a variable selection process could more strongly predict the fate of different borrowings, yielding more information about what causes novel lexical items to succeed or fail. Another comparison which would almost certainly be profitable in this regard is the lexical-
dynamic behavior of novel anglicisms vs. novel slang in German, in order to better tease apart the roles that language of origin and novelty play in this process. A possible factor in an expanded anglicism fate prediction model is inspired by the need vs. prestige distinction with regard to motivation for borrowing, or the 'necessary borrowing'/luxury loan' ideology. If this folk distinction has some real basis, it stands to reason that these types of borrowings might be expected to behave differently in terms of their adoption and adaptation in the German hip hop community. A distributional analysis of the environments of anglicism forms would reveal likely candidates for a German counterpart; anglicisms with close German counterparts would then be compared against those without. Such an analysis stands to illuminate the legitimacy of the necessary borrowing/luxury loan distinction, which, if found, could then be implemented as an additional variable in the model predicting loanword fate.

Additionally, the automated morphological analysis of inflected and derived anglicism forms could open the door for several avenues of inquiry: First, the entropy, a measure of the diversity of word forms for a stem (Gries 2010) could be calculated over different time windows for English-origin word stems, compared to the entropy of native German stems, and implemented as a predictor variable in modeling word frequency change. Second, the distribution of inflected/derived vs. bare stems or stems with English morphology could reveal whether certain stems or classes of stems (parts of speech, or those fulfilling different borrowing needs) are more or less likely to be nativized. An investigation of whether, across stems, there is some agreement or pattern to the sequential occurrence of different types of German affixes on a borrowed word over time could provide a measure of how nativized a word stem is, and serve as an additional factor for this kind of predictive model.

8.5 Contributions of this dissertation

Because of its broad methodological and theoretical orientation, this dissertation contributes to enduring questions of sociohistorical linguistics: first and foremost, it sheds light on the process of linguistic borrowing and provides a novel view of the dynamics of borrowed linguistic material within a globalizing subculture; second, it provides a prototype for the nascent field of large-corpus sociolinguistics; third, it begins to explore the motivations for the linguistic and distributional properties of borrowed linguistic material by exploring attitudes and ideologies.
within the German hip hop community. Finally, this dissertation serves the non-academic community by weighing in on popular metalinguistic debates, detailed in section 1.1 and throughout the dissertation, finding no particular cause for concern in terms of the extent of English influence on German or its potential to somehow disrupt, contribute to the decline of, or pose an existential threat to the German language.
REFERENCES


184


# | Word     | Freq.  | # | Word     | Freq.  | # | Word    | Freq.  
---|----------|--------|---|----------|--------|---|---------|--------
1  | hiphop   | 23883  | 41 | down     | 1522  | 81 | nigga   | 942    
2  | hip      | 23540  | 42 | dissen   | 1456  | 82 | bozz    | 941    
3  | hop      | 22146  | 43 | your     | 1448  | 84 | can     | 930    
4  | track    | 17774  | 44 | bitch    | 941   |    |         |        
5  | tracks   | 11231  | 45 | mic      | 939   |    |         |        
6  | peace    | 7855   | 46 | rhymes   | 942   |    |         |        
7  | rappen   | 7500   | 47 | feature  | 942   |    |         |        
8  | feat     | 6393   | 48 | sampler  | 942   |    |         |        
9  | flow     | 5394   | 49 | mainstream | 942 |    |         |        
10 | style    | 5219   | 50 | snippet  | 942   |    |         |        
11 | mixtape  | 4871   | 51 | mixtapes | 942   |    |         |        
12 | label    | 4713   | 52 | line     | 942   |    |         |        
13 | shit     | 4544   | 53 | producer | 942   |    |         |        
14 | word     | 4064   | 54 | acts     | 942   |    |         |        
15 | gangsta  | 3611   | 55 | hater    | 942   |    |         |        
16 | rappt    | 3608   | 56 | haten    | 942   |    |         |        
17 | battle   | 3299   | 57 | free     | 942   |    |         |        
18 | records  | 3048   | 58 | labels   | 942   |    |         |        
19 | crew     | 2987   | 59 | kiddies  | 942   |    |         |        
20 | hey      | 2601   | 60 | fake     | 942   |    |         |        
21 | dope     | 2292   | 61 | hook     | 942   |    |         |        
22 | sound    | 2285   | 62 | lines    | 942   |    |         |        
23 | wack     | 2225   | 63 | fame     | 942   |    |         |        
24 | music    | 2223   | 64 | disst    | 942   |    |         |        
25 | charts   | 2143   | 65 | burner   | 942   |    |         |        
26 | single   | 2108   | 66 | releases | 942   |    |         |        
27 | lyrics   | 2040   | 67 | young    | 942   |    |         |        
28 | diss     | 1986   | 68 | styles   | 942   |    |         |        
29 | beef     | 1953   | 69 | rule     | 942   |    |         |        
30 | yeah     | 1857   | 70 | hiphopper | 942 |    |         |        
31 | beatz    | 1826   | 71 | jams     | 942   |    |         |        
32 | money    | 1813   | 72 | german   | 942   |    |         |        
33 | close    | 1778   | 73 | oldschool | 942 |    |         |        
34 | features | 1768   | 74 | gerappt  | 942   |    |         |        
35 | life     | 1750   | 75 | hopper   | 942   |    |         |        
36 | edit     | 1750   | 76 | nice     | 942   |    |         |        
37 | release  | 1727   | 77 | gedisst  | 942   |    |         |        
38 | jam      | 1711   | 78 | street   | 942   |    |         |        
39 | love     | 1663   | 79 | who      | 942   |    |         |        
40 | cover    | 1576   | 80 | now      | 942   |    |         |        

APPENDIX A. LIST OF HIGH-FREQUENCY ANGLICISMS IN THE MZEE CORPUS
<table>
<thead>
<tr>
<th>#</th>
<th>Word</th>
<th>Freq.</th>
<th>#</th>
<th>Word</th>
<th>Freq.</th>
<th>#</th>
<th>Word</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>121</td>
<td>blaze</td>
<td>671</td>
<td>161</td>
<td>say</td>
<td>524</td>
<td>201</td>
<td>freestyles</td>
<td>414</td>
</tr>
<tr>
<td>122</td>
<td>day</td>
<td>667</td>
<td>162</td>
<td>homies</td>
<td>524</td>
<td>202</td>
<td>true</td>
<td>408</td>
</tr>
<tr>
<td>123</td>
<td>how</td>
<td>656</td>
<td>163</td>
<td>flash</td>
<td>522</td>
<td>203</td>
<td>rappe</td>
<td>405</td>
</tr>
<tr>
<td>124</td>
<td>deal</td>
<td>645</td>
<td>164</td>
<td>never</td>
<td>515</td>
<td>204</td>
<td>featuring</td>
<td>404</td>
</tr>
<tr>
<td>125</td>
<td>hardcore</td>
<td>638</td>
<td>165</td>
<td>damn</td>
<td>515</td>
<td>205</td>
<td>forever</td>
<td>398</td>
</tr>
<tr>
<td>126</td>
<td>make</td>
<td>631</td>
<td>166</td>
<td>open</td>
<td>511</td>
<td>206</td>
<td>chill</td>
<td>398</td>
</tr>
<tr>
<td>127</td>
<td>thats</td>
<td>623</td>
<td>167</td>
<td>skit</td>
<td>510</td>
<td>207</td>
<td>soundtrack</td>
<td>396</td>
</tr>
<tr>
<td>128</td>
<td>punchlines</td>
<td>613</td>
<td>168</td>
<td>act</td>
<td>510</td>
<td>208</td>
<td>then</td>
<td>390</td>
</tr>
<tr>
<td>129</td>
<td>produced</td>
<td>613</td>
<td>169</td>
<td>here</td>
<td>506</td>
<td>209</td>
<td>dissrack</td>
<td>388</td>
</tr>
<tr>
<td>130</td>
<td>battlerap</td>
<td>608</td>
<td>170</td>
<td>props</td>
<td>493</td>
<td>210</td>
<td>groupies</td>
<td>382</td>
</tr>
<tr>
<td>131</td>
<td>tracklist</td>
<td>603</td>
<td>171</td>
<td>been</td>
<td>493</td>
<td>211</td>
<td>want</td>
<td>380</td>
</tr>
<tr>
<td>132</td>
<td>rapt</td>
<td>600</td>
<td>172</td>
<td>rapmusik</td>
<td>487</td>
<td>212</td>
<td>rip</td>
<td>380</td>
</tr>
<tr>
<td>133</td>
<td>hate</td>
<td>600</td>
<td>173</td>
<td>death</td>
<td>483</td>
<td>213</td>
<td>member</td>
<td>380</td>
</tr>
<tr>
<td>134</td>
<td>right</td>
<td>596</td>
<td>174</td>
<td>biten</td>
<td>478</td>
<td>214</td>
<td>singles</td>
<td>379</td>
</tr>
<tr>
<td>135</td>
<td>there</td>
<td>592</td>
<td>175</td>
<td>crack</td>
<td>477</td>
<td>215</td>
<td>headz</td>
<td>378</td>
</tr>
<tr>
<td>136</td>
<td>geflasht</td>
<td>584</td>
<td>176</td>
<td>flowt</td>
<td>475</td>
<td>216</td>
<td>djing</td>
<td>378</td>
</tr>
<tr>
<td>137</td>
<td>only</td>
<td>582</td>
<td>177</td>
<td>groupie</td>
<td>464</td>
<td>217</td>
<td>burnt</td>
<td>378</td>
</tr>
<tr>
<td>138</td>
<td>week</td>
<td>580</td>
<td>178</td>
<td>deepe</td>
<td>460</td>
<td>218</td>
<td>flasht</td>
<td>375</td>
</tr>
<tr>
<td>139</td>
<td>battlen</td>
<td>580</td>
<td>179</td>
<td>fakers</td>
<td>458</td>
<td>219</td>
<td>need</td>
<td>373</td>
</tr>
<tr>
<td>140</td>
<td>realness</td>
<td>579</td>
<td>180</td>
<td>punchline</td>
<td>451</td>
<td>220</td>
<td>them</td>
<td>372</td>
</tr>
<tr>
<td>141</td>
<td>aint</td>
<td>572</td>
<td>181</td>
<td>sellout</td>
<td>448</td>
<td>221</td>
<td>nigger</td>
<td>370</td>
</tr>
<tr>
<td>142</td>
<td>statement</td>
<td>571</td>
<td>182</td>
<td>weak</td>
<td>445</td>
<td>222</td>
<td>were</td>
<td>369</td>
</tr>
<tr>
<td>143</td>
<td>pimp</td>
<td>570</td>
<td>183</td>
<td>next</td>
<td>445</td>
<td>223</td>
<td>hard</td>
<td>368</td>
</tr>
<tr>
<td>144</td>
<td>power</td>
<td>567</td>
<td>184</td>
<td>straight</td>
<td>441</td>
<td>224</td>
<td>dance</td>
<td>367</td>
</tr>
<tr>
<td>145</td>
<td>god</td>
<td>559</td>
<td>185</td>
<td>weed</td>
<td>436</td>
<td>225</td>
<td>pussy</td>
<td>366</td>
</tr>
<tr>
<td>146</td>
<td>respect</td>
<td>558</td>
<td>186</td>
<td>think</td>
<td>436</td>
<td>226</td>
<td>amirap</td>
<td>364</td>
</tr>
<tr>
<td>147</td>
<td>battles</td>
<td>556</td>
<td>187</td>
<td>had</td>
<td>432</td>
<td>227</td>
<td>afro</td>
<td>363</td>
</tr>
<tr>
<td>148</td>
<td>newcomer</td>
<td>550</td>
<td>188</td>
<td>threat</td>
<td>431</td>
<td>228</td>
<td>eastcoast</td>
<td>360</td>
</tr>
<tr>
<td>149</td>
<td>homie</td>
<td>547</td>
<td>189</td>
<td>non</td>
<td>431</td>
<td>229</td>
<td>lady</td>
<td>358</td>
</tr>
<tr>
<td>150</td>
<td>releasen</td>
<td>544</td>
<td>190</td>
<td>breakdance</td>
<td>430</td>
<td>230</td>
<td>their</td>
<td>356</td>
</tr>
<tr>
<td>151</td>
<td>community</td>
<td>540</td>
<td>191</td>
<td>mcees</td>
<td>428</td>
<td>231</td>
<td>skits</td>
<td>356</td>
</tr>
<tr>
<td>152</td>
<td>classic</td>
<td>539</td>
<td>192</td>
<td>take</td>
<td>425</td>
<td>232</td>
<td>cap</td>
<td>356</td>
</tr>
<tr>
<td>153</td>
<td>player</td>
<td>537</td>
<td>193</td>
<td>crank</td>
<td>425</td>
<td>233</td>
<td>yes</td>
<td>354</td>
</tr>
<tr>
<td>154</td>
<td>come</td>
<td>536</td>
<td>194</td>
<td>exclusive</td>
<td>421</td>
<td>234</td>
<td>blood</td>
<td>353</td>
</tr>
<tr>
<td>155</td>
<td>whack</td>
<td>535</td>
<td>195</td>
<td>where</td>
<td>418</td>
<td>235</td>
<td>cypher</td>
<td>352</td>
</tr>
<tr>
<td>156</td>
<td>baggy</td>
<td>532</td>
<td>196</td>
<td>muzik</td>
<td>418</td>
<td>236</td>
<td>move</td>
<td>351</td>
</tr>
<tr>
<td>157</td>
<td>mod</td>
<td>531</td>
<td>197</td>
<td>hiphops</td>
<td>418</td>
<td>237</td>
<td>board</td>
<td>349</td>
</tr>
<tr>
<td>158</td>
<td>put</td>
<td>529</td>
<td>198</td>
<td>rhyme</td>
<td>417</td>
<td>238</td>
<td>headliners</td>
<td>348</td>
</tr>
<tr>
<td>159</td>
<td>head</td>
<td>529</td>
<td>199</td>
<td>chillen</td>
<td>415</td>
<td>239</td>
<td>project</td>
<td>346</td>
</tr>
<tr>
<td>160</td>
<td>toys</td>
<td>526</td>
<td>200</td>
<td>breaken</td>
<td>415</td>
<td>240</td>
<td>greetz</td>
<td>345</td>
</tr>
<tr>
<td>#</td>
<td>Word</td>
<td>Freq.</td>
<td>#</td>
<td>Word</td>
<td>Freq.</td>
<td>#</td>
<td>Word</td>
<td>Freq.</td>
</tr>
<tr>
<td>----</td>
<td>----------</td>
<td>-------</td>
<td>----</td>
<td>----------</td>
<td>-------</td>
<td>----</td>
<td>----------</td>
<td>-------</td>
</tr>
<tr>
<td>241</td>
<td>well</td>
<td>341</td>
<td>281</td>
<td>record</td>
<td>293</td>
<td>321</td>
<td>scene</td>
<td>261</td>
</tr>
<tr>
<td>242</td>
<td>wanna</td>
<td>341</td>
<td>282</td>
<td>outta</td>
<td>293</td>
<td>322</td>
<td>entertain</td>
<td>261</td>
</tr>
<tr>
<td>243</td>
<td>give</td>
<td>339</td>
<td>283</td>
<td>drums</td>
<td>292</td>
<td>323</td>
<td>freetrack</td>
<td>260</td>
</tr>
<tr>
<td>244</td>
<td>easy</td>
<td>339</td>
<td>284</td>
<td>storytelling</td>
<td>291</td>
<td>324</td>
<td>into</td>
<td>258</td>
</tr>
<tr>
<td>245</td>
<td>signen</td>
<td>337</td>
<td>285</td>
<td>lets</td>
<td>291</td>
<td>325</td>
<td>flowen</td>
<td>258</td>
</tr>
<tr>
<td>246</td>
<td>town</td>
<td>335</td>
<td>286</td>
<td>other</td>
<td>290</td>
<td>326</td>
<td>much</td>
<td>257</td>
</tr>
<tr>
<td>247</td>
<td>english</td>
<td>333</td>
<td>287</td>
<td>full</td>
<td>289</td>
<td>327</td>
<td>kings</td>
<td>256</td>
</tr>
<tr>
<td>248</td>
<td>disses</td>
<td>333</td>
<td>288</td>
<td>sounds</td>
<td>288</td>
<td>328</td>
<td>many</td>
<td>255</td>
</tr>
<tr>
<td>249</td>
<td>breaker</td>
<td>332</td>
<td>289</td>
<td>yall</td>
<td>287</td>
<td>329</td>
<td>fire</td>
<td>255</td>
</tr>
<tr>
<td>250</td>
<td>than</td>
<td>331</td>
<td>290</td>
<td>hooks</td>
<td>287</td>
<td>330</td>
<td>collabo</td>
<td>255</td>
</tr>
<tr>
<td>251</td>
<td>proof</td>
<td>331</td>
<td>291</td>
<td>gonna</td>
<td>286</td>
<td>331</td>
<td>chart</td>
<td>255</td>
</tr>
<tr>
<td>252</td>
<td>cant</td>
<td>331</td>
<td>292</td>
<td>trackz</td>
<td>285</td>
<td>332</td>
<td>our</td>
<td>254</td>
</tr>
<tr>
<td>253</td>
<td>chronic</td>
<td>330</td>
<td>293</td>
<td>takeover</td>
<td>285</td>
<td>333</td>
<td>copies</td>
<td>254</td>
</tr>
<tr>
<td>254</td>
<td>push</td>
<td>328</td>
<td>294</td>
<td>really</td>
<td>284</td>
<td>334</td>
<td>hoes</td>
<td>253</td>
</tr>
<tr>
<td>255</td>
<td>flip</td>
<td>328</td>
<td>295</td>
<td>takeover</td>
<td>284</td>
<td>335</td>
<td>freestyler</td>
<td>253</td>
</tr>
<tr>
<td>256</td>
<td>deepen</td>
<td>324</td>
<td>296</td>
<td>hangout</td>
<td>283</td>
<td>336</td>
<td>whats</td>
<td>251</td>
</tr>
<tr>
<td>257</td>
<td>crazy</td>
<td>323</td>
<td>297</td>
<td>gedropt</td>
<td>283</td>
<td>337</td>
<td>sentence</td>
<td>250</td>
</tr>
<tr>
<td>258</td>
<td>even</td>
<td>322</td>
<td>298</td>
<td>hatet</td>
<td>284</td>
<td>335</td>
<td>freestyler</td>
<td>253</td>
</tr>
<tr>
<td>259</td>
<td>peaz</td>
<td>321</td>
<td>299</td>
<td>did</td>
<td>282</td>
<td>339</td>
<td>feeling</td>
<td>249</td>
</tr>
<tr>
<td>260</td>
<td>holla</td>
<td>321</td>
<td>300</td>
<td>search</td>
<td>281</td>
<td>340</td>
<td>account</td>
<td>249</td>
</tr>
<tr>
<td>261</td>
<td>fun</td>
<td>318</td>
<td>301</td>
<td>omg</td>
<td>279</td>
<td>341</td>
<td>shorty</td>
<td>248</td>
</tr>
<tr>
<td>262</td>
<td>peazen</td>
<td>317</td>
<td>302</td>
<td>phat</td>
<td>276</td>
<td>342</td>
<td>lyricist</td>
<td>248</td>
</tr>
<tr>
<td>263</td>
<td>outro</td>
<td>317</td>
<td>303</td>
<td>rulez</td>
<td>273</td>
<td>343</td>
<td>feel</td>
<td>248</td>
</tr>
<tr>
<td>264</td>
<td>poser</td>
<td>313</td>
<td>304</td>
<td>flashy</td>
<td>273</td>
<td>344</td>
<td>dude</td>
<td>248</td>
</tr>
<tr>
<td>265</td>
<td>aight</td>
<td>309</td>
<td>305</td>
<td>releast</td>
<td>272</td>
<td>345</td>
<td>again</td>
<td>248</td>
</tr>
<tr>
<td>266</td>
<td>which</td>
<td>308</td>
<td>306</td>
<td>connections</td>
<td>272</td>
<td>346</td>
<td>years</td>
<td>247</td>
</tr>
<tr>
<td>267</td>
<td>double</td>
<td>308</td>
<td>307</td>
<td>nigga</td>
<td>271</td>
<td>347</td>
<td>please</td>
<td>247</td>
</tr>
<tr>
<td>268</td>
<td>bitet</td>
<td>308</td>
<td>308</td>
<td>infamous</td>
<td>271</td>
<td>348</td>
<td>date</td>
<td>247</td>
</tr>
<tr>
<td>269</td>
<td>baggies</td>
<td>307</td>
<td>309</td>
<td>better</td>
<td>271</td>
<td>349</td>
<td>chillig</td>
<td>247</td>
</tr>
<tr>
<td>270</td>
<td>stay</td>
<td>306</td>
<td>310</td>
<td>whatever</td>
<td>270</td>
<td>350</td>
<td>jep</td>
<td>246</td>
</tr>
<tr>
<td>271</td>
<td>members</td>
<td>306</td>
<td>311</td>
<td>watch</td>
<td>269</td>
<td>351</td>
<td>thing</td>
<td>245</td>
</tr>
<tr>
<td>272</td>
<td>tighte</td>
<td>305</td>
<td>312</td>
<td>american</td>
<td>269</td>
<td>352</td>
<td>something</td>
<td>244</td>
</tr>
<tr>
<td>273</td>
<td>hiphoper</td>
<td>304</td>
<td>313</td>
<td>face</td>
<td>267</td>
<td>353</td>
<td>caps</td>
<td>244</td>
</tr>
<tr>
<td>274</td>
<td>coast</td>
<td>300</td>
<td>314</td>
<td>coming</td>
<td>267</td>
<td>354</td>
<td>hustler</td>
<td>243</td>
</tr>
<tr>
<td>275</td>
<td>droppen</td>
<td>299</td>
<td>315</td>
<td>year</td>
<td>266</td>
<td>355</td>
<td>gangsterrap</td>
<td>243</td>
</tr>
<tr>
<td>276</td>
<td>night</td>
<td>298</td>
<td>316</td>
<td>shot</td>
<td>266</td>
<td>356</td>
<td>drop</td>
<td>243</td>
</tr>
<tr>
<td>277</td>
<td>booklet</td>
<td>297</td>
<td>317</td>
<td>kill</td>
<td>264</td>
<td>357</td>
<td>gangstarap</td>
<td>242</td>
</tr>
<tr>
<td>278</td>
<td>look</td>
<td>295</td>
<td>318</td>
<td>mom</td>
<td>262</td>
<td>358</td>
<td>till</td>
<td>241</td>
</tr>
<tr>
<td>279</td>
<td>would</td>
<td>294</td>
<td>319</td>
<td>goes</td>
<td>262</td>
<td>359</td>
<td>tell</td>
<td>241</td>
</tr>
<tr>
<td>280</td>
<td>because</td>
<td>294</td>
<td>320</td>
<td>before</td>
<td>262</td>
<td>360</td>
<td>independent</td>
<td>241</td>
</tr>
</tbody>
</table>
### APPENDIX A, CONT'D.

<table>
<thead>
<tr>
<th>#</th>
<th>Word</th>
<th>Freq.</th>
<th>#</th>
<th>Word</th>
<th>Freq.</th>
<th>#</th>
<th>Word</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>361</td>
<td>gone</td>
<td>241</td>
<td>361</td>
<td>sooooo</td>
<td>202</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>362</td>
<td>army</td>
<td>241</td>
<td>362</td>
<td>playa</td>
<td>202</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>363</td>
<td>honk</td>
<td>240</td>
<td>363</td>
<td>newsletter</td>
<td>202</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>364</td>
<td>conscious</td>
<td>240</td>
<td>364</td>
<td>heard</td>
<td>202</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>365</td>
<td>blade</td>
<td>239</td>
<td>365</td>
<td>female</td>
<td>202</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>366</td>
<td>dis</td>
<td>238</td>
<td>366</td>
<td>yourself</td>
<td>201</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>367</td>
<td>green</td>
<td>237</td>
<td>367</td>
<td>men</td>
<td>201</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>368</td>
<td>biter</td>
<td>237</td>
<td>368</td>
<td>havoc</td>
<td>201</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>369</td>
<td>state</td>
<td>236</td>
<td>369</td>
<td>ganxta</td>
<td>201</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>370</td>
<td>hops</td>
<td>236</td>
<td>370</td>
<td>kiff</td>
<td>200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>371</td>
<td>truth</td>
<td>235</td>
<td>371</td>
<td>illz</td>
<td>200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>372</td>
<td>rules</td>
<td>235</td>
<td>372</td>
<td>fab</td>
<td>200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>373</td>
<td>classics</td>
<td>233</td>
<td>373</td>
<td>remixe</td>
<td>199</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>374</td>
<td>through</td>
<td>232</td>
<td>374</td>
<td>presents</td>
<td>199</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>375</td>
<td>side</td>
<td>232</td>
<td>375</td>
<td>murda</td>
<td>199</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>376</td>
<td>roll</td>
<td>232</td>
<td>376</td>
<td>born</td>
<td>199</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>377</td>
<td>kidz</td>
<td>232</td>
<td>377</td>
<td>orginal</td>
<td>198</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>378</td>
<td>gee</td>
<td>232</td>
<td>378</td>
<td>chorus</td>
<td>198</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>379</td>
<td>awards</td>
<td>232</td>
<td>379</td>
<td>supreme</td>
<td>197</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>380</td>
<td>lion</td>
<td>231</td>
<td>380</td>
<td>skyline</td>
<td>197</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>381</td>
<td>grime</td>
<td>231</td>
<td>381</td>
<td>enter</td>
<td>197</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>382</td>
<td>around</td>
<td>231</td>
<td>382</td>
<td>greatest</td>
<td>195</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>383</td>
<td>statements</td>
<td>230</td>
<td>383</td>
<td>fella</td>
<td>195</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>384</td>
<td>gig</td>
<td>230</td>
<td>384</td>
<td>vibe</td>
<td>194</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>385</td>
<td>era</td>
<td>230</td>
<td>385</td>
<td>could</td>
<td>194</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>386</td>
<td>bye</td>
<td>229</td>
<td>386</td>
<td>ride</td>
<td>193</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>387</td>
<td>attack</td>
<td>229</td>
<td>387</td>
<td>blue</td>
<td>193</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>388</td>
<td>scratches</td>
<td>228</td>
<td>388</td>
<td>against</td>
<td>193</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>389</td>
<td>five</td>
<td>225</td>
<td>389</td>
<td>french</td>
<td>192</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>390</td>
<td>work</td>
<td>224</td>
<td>390</td>
<td>earth</td>
<td>192</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>391</td>
<td>poster</td>
<td>224</td>
<td>391</td>
<td>crowd</td>
<td>192</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>392</td>
<td>lad</td>
<td>224</td>
<td>392</td>
<td>groove</td>
<td>191</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>393</td>
<td>ghostwriter</td>
<td>224</td>
<td>393</td>
<td>ones</td>
<td>190</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>394</td>
<td>going</td>
<td>223</td>
<td>394</td>
<td>every</td>
<td>190</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>395</td>
<td>ready</td>
<td>222</td>
<td>395</td>
<td>whole</td>
<td>189</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>396</td>
<td>doubletime</td>
<td>222</td>
<td>396</td>
<td>says</td>
<td>189</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>397</td>
<td>dancehall</td>
<td>222</td>
<td>397</td>
<td>probes</td>
<td>189</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>398</td>
<td>dark</td>
<td>221</td>
<td>398</td>
<td>hobo</td>
<td>188</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>399</td>
<td>call</td>
<td>221</td>
<td>399</td>
<td>won</td>
<td>187</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>superstar</td>
<td>220</td>
<td>400</td>
<td>things</td>
<td>187</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# APPENDIX A, CONT'D.

<table>
<thead>
<tr>
<th>#</th>
<th>Word</th>
<th>Freq.</th>
<th>#</th>
<th>Word</th>
<th>Freq.</th>
<th>#</th>
<th>Word</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>481</td>
<td>moves</td>
<td>187</td>
<td>521</td>
<td>kick</td>
<td>173</td>
<td>561</td>
<td>skill</td>
<td>158</td>
</tr>
<tr>
<td>482</td>
<td>gotta</td>
<td>187</td>
<td>522</td>
<td>imo</td>
<td>173</td>
<td>562</td>
<td>may</td>
<td>158</td>
</tr>
<tr>
<td>483</td>
<td>rough</td>
<td>186</td>
<td>523</td>
<td>while</td>
<td>172</td>
<td>563</td>
<td>friday</td>
<td>158</td>
</tr>
<tr>
<td>484</td>
<td>walk</td>
<td>185</td>
<td>524</td>
<td>seperate</td>
<td>172</td>
<td>564</td>
<td>anthem</td>
<td>158</td>
</tr>
<tr>
<td>485</td>
<td>tighter</td>
<td>185</td>
<td>525</td>
<td>screw</td>
<td>172</td>
<td>565</td>
<td>shut</td>
<td>157</td>
</tr>
<tr>
<td>486</td>
<td>raper</td>
<td>185</td>
<td>526</td>
<td>since</td>
<td>170</td>
<td>566</td>
<td>featured</td>
<td>157</td>
</tr>
<tr>
<td>487</td>
<td>fable</td>
<td>184</td>
<td>527</td>
<td>flown</td>
<td>170</td>
<td>567</td>
<td>doing</td>
<td>156</td>
</tr>
<tr>
<td>488</td>
<td>any</td>
<td>184</td>
<td>528</td>
<td>signing</td>
<td>169</td>
<td>568</td>
<td>propz</td>
<td>155</td>
</tr>
<tr>
<td>489</td>
<td>hoe</td>
<td>183</td>
<td>529</td>
<td>reg</td>
<td>169</td>
<td>569</td>
<td>featuren</td>
<td>155</td>
</tr>
<tr>
<td>490</td>
<td>bless</td>
<td>183</td>
<td>530</td>
<td>everything</td>
<td>169</td>
<td>570</td>
<td>clips</td>
<td>155</td>
</tr>
<tr>
<td>491</td>
<td>thought</td>
<td>182</td>
<td>531</td>
<td>sparring</td>
<td>167</td>
<td>571</td>
<td>ups</td>
<td>154</td>
</tr>
<tr>
<td>492</td>
<td>nothing</td>
<td>182</td>
<td>532</td>
<td>samplen</td>
<td>167</td>
<td>572</td>
<td>those</td>
<td>154</td>
</tr>
<tr>
<td>493</td>
<td>control</td>
<td>182</td>
<td>533</td>
<td>america</td>
<td>167</td>
<td>573</td>
<td>happy</td>
<td>154</td>
</tr>
<tr>
<td>494</td>
<td>tighten</td>
<td>181</td>
<td>534</td>
<td>peanuts</td>
<td>166</td>
<td>574</td>
<td>swag</td>
<td>153</td>
</tr>
<tr>
<td>495</td>
<td>screwed</td>
<td>181</td>
<td>535</td>
<td>games</td>
<td>166</td>
<td>575</td>
<td>store</td>
<td>153</td>
</tr>
<tr>
<td>496</td>
<td>drive</td>
<td>181</td>
<td>536</td>
<td>flyer</td>
<td>166</td>
<td>576</td>
<td>indie</td>
<td>153</td>
</tr>
<tr>
<td>497</td>
<td>kiddie</td>
<td>180</td>
<td>537</td>
<td>newschool</td>
<td>165</td>
<td>577</td>
<td>flop</td>
<td>153</td>
</tr>
<tr>
<td>498</td>
<td>gangstas</td>
<td>180</td>
<td>538</td>
<td>hear</td>
<td>165</td>
<td>578</td>
<td>fave</td>
<td>153</td>
</tr>
<tr>
<td>499</td>
<td>rappst</td>
<td>179</td>
<td>539</td>
<td>scratch</td>
<td>164</td>
<td>579</td>
<td>cold</td>
<td>153</td>
</tr>
<tr>
<td>500</td>
<td>ohh</td>
<td>179</td>
<td>540</td>
<td>must</td>
<td>164</td>
<td>580</td>
<td>change</td>
<td>153</td>
</tr>
<tr>
<td>501</td>
<td>movement</td>
<td>179</td>
<td>541</td>
<td>fight</td>
<td>164</td>
<td>581</td>
<td>body</td>
<td>153</td>
</tr>
<tr>
<td>502</td>
<td>closed</td>
<td>179</td>
<td>542</td>
<td>closed</td>
<td>163</td>
<td>582</td>
<td>yep</td>
<td>152</td>
</tr>
<tr>
<td>503</td>
<td>area</td>
<td>179</td>
<td>543</td>
<td>welcome</td>
<td>162</td>
<td>583</td>
<td>tread</td>
<td>152</td>
</tr>
<tr>
<td>504</td>
<td>andres</td>
<td>179</td>
<td>544</td>
<td>spice</td>
<td>162</td>
<td>584</td>
<td>duck</td>
<td>152</td>
</tr>
<tr>
<td>505</td>
<td>ciao</td>
<td>178</td>
<td>545</td>
<td>feats</td>
<td>162</td>
<td>585</td>
<td>used</td>
<td>151</td>
</tr>
<tr>
<td>506</td>
<td>another</td>
<td>178</td>
<td>546</td>
<td>done</td>
<td>162</td>
<td>586</td>
<td>trailer</td>
<td>151</td>
</tr>
<tr>
<td>507</td>
<td>trax</td>
<td>177</td>
<td>547</td>
<td>youre</td>
<td>161</td>
<td>587</td>
<td>threadersteller</td>
<td>151</td>
</tr>
<tr>
<td>508</td>
<td>jump</td>
<td>177</td>
<td>548</td>
<td>swagger</td>
<td>161</td>
<td>588</td>
<td>large</td>
<td>151</td>
</tr>
<tr>
<td>509</td>
<td>funny</td>
<td>177</td>
<td>549</td>
<td>scratchen</td>
<td>161</td>
<td>589</td>
<td>motherer</td>
<td>150</td>
</tr>
<tr>
<td>510</td>
<td>far</td>
<td>177</td>
<td>550</td>
<td>punches</td>
<td>161</td>
<td>590</td>
<td>gothic</td>
<td>150</td>
</tr>
<tr>
<td>511</td>
<td>evidence</td>
<td>177</td>
<td>551</td>
<td>living</td>
<td>161</td>
<td>591</td>
<td>self</td>
<td>149</td>
</tr>
<tr>
<td>512</td>
<td>chillt</td>
<td>177</td>
<td>552</td>
<td>later</td>
<td>161</td>
<td>592</td>
<td>mami</td>
<td>149</td>
</tr>
<tr>
<td>513</td>
<td>writing</td>
<td>175</td>
<td>553</td>
<td>hes</td>
<td>161</td>
<td>593</td>
<td>freeway</td>
<td>149</td>
</tr>
<tr>
<td>514</td>
<td>stones</td>
<td>175</td>
<td>554</td>
<td>doubt</td>
<td>161</td>
<td>594</td>
<td>credibility</td>
<td>149</td>
</tr>
<tr>
<td>515</td>
<td>equipment</td>
<td>175</td>
<td>555</td>
<td>told</td>
<td>160</td>
<td>595</td>
<td>comes</td>
<td>149</td>
</tr>
<tr>
<td>516</td>
<td>always</td>
<td>175</td>
<td>556</td>
<td>spitten</td>
<td>160</td>
<td>596</td>
<td>backstage</td>
<td>149</td>
</tr>
<tr>
<td>517</td>
<td>mcing</td>
<td>174</td>
<td>557</td>
<td>nobody</td>
<td>159</td>
<td>597</td>
<td>rebel</td>
<td>148</td>
</tr>
<tr>
<td>518</td>
<td>ladies</td>
<td>174</td>
<td>558</td>
<td>lif</td>
<td>159</td>
<td>598</td>
<td>purple</td>
<td>148</td>
</tr>
<tr>
<td>519</td>
<td>currency</td>
<td>174</td>
<td>559</td>
<td>history</td>
<td>159</td>
<td>599</td>
<td>loser</td>
<td>148</td>
</tr>
<tr>
<td>520</td>
<td>being</td>
<td>174</td>
<td>560</td>
<td>dealer</td>
<td>159</td>
<td>600</td>
<td>aso</td>
<td>148</td>
</tr>
<tr>
<td>#</td>
<td>Word</td>
<td>Freq.</td>
<td>#</td>
<td>Word</td>
<td>Freq.</td>
<td>#</td>
<td>Word</td>
<td>Freq.</td>
</tr>
<tr>
<td>---</td>
<td>------------</td>
<td>-------</td>
<td>---</td>
<td>------------</td>
<td>-------</td>
<td>---</td>
<td>------------</td>
<td>-------</td>
</tr>
<tr>
<td>601</td>
<td>releasedate</td>
<td>147</td>
<td>643</td>
<td>myself</td>
<td>135</td>
<td>685</td>
<td>child</td>
<td>128</td>
</tr>
<tr>
<td>602</td>
<td>livin</td>
<td>147</td>
<td>644</td>
<td>maybe</td>
<td>135</td>
<td>686</td>
<td>vinyls</td>
<td>127</td>
</tr>
<tr>
<td>603</td>
<td>legends</td>
<td>146</td>
<td>645</td>
<td>culture</td>
<td>135</td>
<td>687</td>
<td>quo</td>
<td>127</td>
</tr>
<tr>
<td>604</td>
<td>great</td>
<td>146</td>
<td>646</td>
<td>worst</td>
<td>134</td>
<td>688</td>
<td>killt</td>
<td>127</td>
</tr>
<tr>
<td>605</td>
<td>encore</td>
<td>146</td>
<td>647</td>
<td>ragga</td>
<td>134</td>
<td>689</td>
<td>food</td>
<td>127</td>
</tr>
<tr>
<td>606</td>
<td>weeks</td>
<td>145</td>
<td>648</td>
<td>queen</td>
<td>134</td>
<td>690</td>
<td>comment</td>
<td>127</td>
</tr>
<tr>
<td>607</td>
<td>friends</td>
<td>145</td>
<td>649</td>
<td>heart</td>
<td>134</td>
<td>691</td>
<td>both</td>
<td>127</td>
</tr>
<tr>
<td>608</td>
<td>fanbase</td>
<td>145</td>
<td>650</td>
<td>gclosed</td>
<td>134</td>
<td>692</td>
<td>bloods</td>
<td>127</td>
</tr>
<tr>
<td>609</td>
<td>background</td>
<td>145</td>
<td>651</td>
<td>stylez</td>
<td>133</td>
<td>693</td>
<td>sprayer</td>
<td>126</td>
</tr>
<tr>
<td>610</td>
<td>everybody</td>
<td>144</td>
<td>652</td>
<td>sticky</td>
<td>133</td>
<td>694</td>
<td>space</td>
<td>126</td>
</tr>
<tr>
<td>611</td>
<td>teenies</td>
<td>143</td>
<td>653</td>
<td>shake</td>
<td>133</td>
<td>695</td>
<td>raver</td>
<td>126</td>
</tr>
<tr>
<td>612</td>
<td>should</td>
<td>143</td>
<td>654</td>
<td>clean</td>
<td>133</td>
<td>696</td>
<td>oldskool</td>
<td>126</td>
</tr>
<tr>
<td>613</td>
<td>rapen</td>
<td>143</td>
<td>655</td>
<td>stream</td>
<td>132</td>
<td>697</td>
<td>late</td>
<td>126</td>
</tr>
<tr>
<td>614</td>
<td>horrorcore</td>
<td>143</td>
<td>656</td>
<td>reviews</td>
<td>132</td>
<td>698</td>
<td>eyes</td>
<td>126</td>
</tr>
<tr>
<td>615</td>
<td>alive</td>
<td>142</td>
<td>657</td>
<td>kollabos</td>
<td>132</td>
<td>699</td>
<td>ear</td>
<td>126</td>
</tr>
<tr>
<td>616</td>
<td>wear</td>
<td>141</td>
<td>658</td>
<td>keys</td>
<td>132</td>
<td>700</td>
<td>collabos</td>
<td>126</td>
</tr>
<tr>
<td>617</td>
<td>tryin</td>
<td>141</td>
<td>659</td>
<td>insider</td>
<td>132</td>
<td>701</td>
<td>turntables</td>
<td>125</td>
</tr>
<tr>
<td>618</td>
<td>second</td>
<td>141</td>
<td>660</td>
<td>gesigned</td>
<td>132</td>
<td>702</td>
<td>reality</td>
<td>125</td>
</tr>
<tr>
<td>619</td>
<td>return</td>
<td>141</td>
<td>661</td>
<td>chopped</td>
<td>132</td>
<td>703</td>
<td>rapperin</td>
<td>125</td>
</tr>
<tr>
<td>620</td>
<td>left</td>
<td>141</td>
<td>662</td>
<td>wrong</td>
<td>131</td>
<td>704</td>
<td>movie</td>
<td>125</td>
</tr>
<tr>
<td>621</td>
<td>cya</td>
<td>141</td>
<td>663</td>
<td>whos</td>
<td>131</td>
<td>705</td>
<td>each</td>
<td>125</td>
</tr>
<tr>
<td>622</td>
<td>brain</td>
<td>141</td>
<td>664</td>
<td>promoten</td>
<td>131</td>
<td>706</td>
<td>rapp</td>
<td>124</td>
</tr>
<tr>
<td>623</td>
<td>rugged</td>
<td>140</td>
<td>665</td>
<td>knowledge</td>
<td>131</td>
<td>707</td>
<td>performance</td>
<td>124</td>
</tr>
<tr>
<td>624</td>
<td>knock</td>
<td>140</td>
<td>666</td>
<td>doper</td>
<td>131</td>
<td>708</td>
<td>ohhh</td>
<td>124</td>
</tr>
<tr>
<td>625</td>
<td>highlight</td>
<td>140</td>
<td>667</td>
<td>contest</td>
<td>131</td>
<td>709</td>
<td>burn</td>
<td>124</td>
</tr>
<tr>
<td>626</td>
<td>crips</td>
<td>140</td>
<td>668</td>
<td>believe</td>
<td>131</td>
<td>710</td>
<td>hehehe</td>
<td>123</td>
</tr>
<tr>
<td>627</td>
<td>supporter</td>
<td>139</td>
<td>669</td>
<td>wicked</td>
<td>130</td>
<td>711</td>
<td>fam</td>
<td>123</td>
</tr>
<tr>
<td>628</td>
<td>came</td>
<td>139</td>
<td>670</td>
<td>uncool</td>
<td>130</td>
<td>712</td>
<td>crossover</td>
<td>123</td>
</tr>
<tr>
<td>629</td>
<td>size</td>
<td>138</td>
<td>671</td>
<td>together</td>
<td>130</td>
<td>713</td>
<td>backpack</td>
<td>123</td>
</tr>
<tr>
<td>630</td>
<td>sit</td>
<td>138</td>
<td>672</td>
<td>sing</td>
<td>130</td>
<td>714</td>
<td>away</td>
<td>123</td>
</tr>
<tr>
<td>631</td>
<td>mcee</td>
<td>138</td>
<td>673</td>
<td>rult</td>
<td>130</td>
<td>715</td>
<td>vid</td>
<td>122</td>
</tr>
<tr>
<td>632</td>
<td>brothath</td>
<td>138</td>
<td>674</td>
<td>light</td>
<td>130</td>
<td>716</td>
<td>slow</td>
<td>122</td>
</tr>
<tr>
<td>633</td>
<td>without</td>
<td>137</td>
<td>675</td>
<td>fresher</td>
<td>130</td>
<td>717</td>
<td>reason</td>
<td>122</td>
</tr>
<tr>
<td>634</td>
<td>sales</td>
<td>137</td>
<td>676</td>
<td>disse</td>
<td>130</td>
<td>718</td>
<td>geburnt</td>
<td>122</td>
</tr>
<tr>
<td>635</td>
<td>jungz</td>
<td>137</td>
<td>677</td>
<td>kiss</td>
<td>129</td>
<td>719</td>
<td>few</td>
<td>122</td>
</tr>
<tr>
<td>636</td>
<td>hmmmm</td>
<td>137</td>
<td>678</td>
<td>trendsetter</td>
<td>128</td>
<td>720</td>
<td>devil</td>
<td>122</td>
</tr>
<tr>
<td>637</td>
<td>remember</td>
<td>136</td>
<td>679</td>
<td>rapstyle</td>
<td>128</td>
<td>721</td>
<td>darn</td>
<td>122</td>
</tr>
<tr>
<td>638</td>
<td>production</td>
<td>136</td>
<td>680</td>
<td>producing</td>
<td>128</td>
<td>722</td>
<td>angels</td>
<td>122</td>
</tr>
<tr>
<td>639</td>
<td>twisted</td>
<td>135</td>
<td>681</td>
<td>gigs</td>
<td>128</td>
<td>723</td>
<td>vote</td>
<td>121</td>
</tr>
<tr>
<td>640</td>
<td>suckt</td>
<td>135</td>
<td>682</td>
<td>gesamplet</td>
<td>128</td>
<td>724</td>
<td>president</td>
<td>121</td>
</tr>
<tr>
<td>641</td>
<td>preview</td>
<td>135</td>
<td>683</td>
<td>competition</td>
<td>128</td>
<td>725</td>
<td>guest</td>
<td>121</td>
</tr>
<tr>
<td>642</td>
<td>official</td>
<td>135</td>
<td>684</td>
<td>comedy</td>
<td>128</td>
<td>726</td>
<td>distributionz</td>
<td>121</td>
</tr>
</tbody>
</table>
### APPENDIX A, CONT'D.

<table>
<thead>
<tr>
<th>#</th>
<th>Word</th>
<th>Freq.</th>
<th>#</th>
<th>Word</th>
<th>Freq.</th>
<th>#</th>
<th>Word</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>727</td>
<td>broke</td>
<td>121</td>
<td>769</td>
<td>jungle</td>
<td>113</td>
<td>811</td>
<td>asia</td>
<td>106</td>
</tr>
<tr>
<td>728</td>
<td>boo</td>
<td>121</td>
<td>770</td>
<td>guess</td>
<td>113</td>
<td>812</td>
<td>trash</td>
<td>105</td>
</tr>
<tr>
<td>729</td>
<td>working</td>
<td>120</td>
<td>771</td>
<td>voting</td>
<td>112</td>
<td>813</td>
<td>represent</td>
<td>105</td>
</tr>
<tr>
<td>730</td>
<td>place</td>
<td>120</td>
<td>772</td>
<td>sucks</td>
<td>112</td>
<td>814</td>
<td>remixes</td>
<td>105</td>
</tr>
<tr>
<td>731</td>
<td>hung</td>
<td>120</td>
<td>773</td>
<td>mother</td>
<td>112</td>
<td>815</td>
<td>homes</td>
<td>105</td>
</tr>
<tr>
<td>732</td>
<td>gechilt</td>
<td>120</td>
<td>774</td>
<td>loool</td>
<td>112</td>
<td>816</td>
<td>whoo</td>
<td>104</td>
</tr>
<tr>
<td>733</td>
<td>award</td>
<td>120</td>
<td>775</td>
<td>lights</td>
<td>112</td>
<td>817</td>
<td>shoot</td>
<td>104</td>
</tr>
<tr>
<td>734</td>
<td>tragedy</td>
<td>119</td>
<td>776</td>
<td>drugs</td>
<td>112</td>
<td>818</td>
<td>kiddis</td>
<td>104</td>
</tr>
<tr>
<td>735</td>
<td>sure</td>
<td>119</td>
<td>777</td>
<td>went</td>
<td>111</td>
<td>819</td>
<td>getting</td>
<td>104</td>
</tr>
<tr>
<td>736</td>
<td>struggle</td>
<td>119</td>
<td>778</td>
<td>ruled</td>
<td>111</td>
<td>820</td>
<td>flowmässig</td>
<td>104</td>
</tr>
<tr>
<td>737</td>
<td>loop</td>
<td>119</td>
<td>779</td>
<td>manchma</td>
<td>111</td>
<td>821</td>
<td>battletexte</td>
<td>104</td>
</tr>
<tr>
<td>738</td>
<td>deeper</td>
<td>119</td>
<td>780</td>
<td>killen</td>
<td>111</td>
<td>822</td>
<td>ask</td>
<td>104</td>
</tr>
<tr>
<td>739</td>
<td>talking</td>
<td>118</td>
<td>781</td>
<td>ether</td>
<td>111</td>
<td>823</td>
<td>superstars</td>
<td>103</td>
</tr>
<tr>
<td>740</td>
<td>sight</td>
<td>118</td>
<td>782</td>
<td>emcees</td>
<td>111</td>
<td>824</td>
<td>hoppenn</td>
<td>103</td>
</tr>
<tr>
<td>741</td>
<td>help</td>
<td>118</td>
<td>783</td>
<td>writen</td>
<td>110</td>
<td>825</td>
<td>glass</td>
<td>103</td>
</tr>
<tr>
<td>742</td>
<td>guy</td>
<td>118</td>
<td>784</td>
<td>mailorder</td>
<td>110</td>
<td>826</td>
<td>demotape</td>
<td>103</td>
</tr>
<tr>
<td>743</td>
<td>cops</td>
<td>118</td>
<td>785</td>
<td>gedisse</td>
<td>110</td>
<td>827</td>
<td>perfect</td>
<td>102</td>
</tr>
<tr>
<td>744</td>
<td>triple</td>
<td>117</td>
<td>786</td>
<td>busdriver</td>
<td>110</td>
<td>828</td>
<td>past</td>
<td>102</td>
</tr>
<tr>
<td>745</td>
<td>spit</td>
<td>117</td>
<td>787</td>
<td>bet</td>
<td>110</td>
<td>829</td>
<td>oberhammer</td>
<td>102</td>
</tr>
<tr>
<td>746</td>
<td>shops</td>
<td>117</td>
<td>788</td>
<td>stoned</td>
<td>109</td>
<td>830</td>
<td>joke</td>
<td>102</td>
</tr>
<tr>
<td>747</td>
<td>number</td>
<td>117</td>
<td>789</td>
<td>split</td>
<td>109</td>
<td>831</td>
<td>dent</td>
<td>102</td>
</tr>
<tr>
<td>748</td>
<td>thang</td>
<td>116</td>
<td>790</td>
<td>sniper</td>
<td>109</td>
<td>832</td>
<td>country</td>
<td>102</td>
</tr>
<tr>
<td>749</td>
<td>snare</td>
<td>116</td>
<td>791</td>
<td>representen</td>
<td>109</td>
<td>833</td>
<td>casting</td>
<td>102</td>
</tr>
<tr>
<td>750</td>
<td>piece</td>
<td>116</td>
<td>792</td>
<td>peez</td>
<td>109</td>
<td>834</td>
<td>bouncen</td>
<td>102</td>
</tr>
<tr>
<td>751</td>
<td>known</td>
<td>116</td>
<td>793</td>
<td>outfit</td>
<td>109</td>
<td>835</td>
<td>basscrew</td>
<td>102</td>
</tr>
<tr>
<td>752</td>
<td>gehyped</td>
<td>116</td>
<td>794</td>
<td>himself</td>
<td>109</td>
<td>836</td>
<td>bassboxx</td>
<td>102</td>
</tr>
<tr>
<td>753</td>
<td>evil</td>
<td>116</td>
<td>795</td>
<td>event</td>
<td>109</td>
<td>837</td>
<td>until</td>
<td>101</td>
</tr>
<tr>
<td>754</td>
<td>deepes</td>
<td>116</td>
<td>796</td>
<td>breezer</td>
<td>109</td>
<td>838</td>
<td>smut</td>
<td>101</td>
</tr>
<tr>
<td>755</td>
<td>words</td>
<td>115</td>
<td>797</td>
<td>micro</td>
<td>108</td>
<td>839</td>
<td>freshen</td>
<td>101</td>
</tr>
<tr>
<td>756</td>
<td>started</td>
<td>115</td>
<td>798</td>
<td>copy</td>
<td>108</td>
<td>840</td>
<td>dreams</td>
<td>101</td>
</tr>
<tr>
<td>757</td>
<td>signed</td>
<td>115</td>
<td>799</td>
<td>central</td>
<td>108</td>
<td>841</td>
<td>coolste</td>
<td>101</td>
</tr>
<tr>
<td>758</td>
<td>understand</td>
<td>114</td>
<td>800</td>
<td>behind</td>
<td>108</td>
<td>842</td>
<td>age</td>
<td>101</td>
</tr>
<tr>
<td>759</td>
<td>thanks</td>
<td>114</td>
<td>801</td>
<td>beefs</td>
<td>108</td>
<td>843</td>
<td>yassir</td>
<td>100</td>
</tr>
<tr>
<td>760</td>
<td>save</td>
<td>114</td>
<td>802</td>
<td>view</td>
<td>107</td>
<td>844</td>
<td>trying</td>
<td>100</td>
</tr>
<tr>
<td>761</td>
<td>popular</td>
<td>114</td>
<td>803</td>
<td>dad</td>
<td>107</td>
<td>845</td>
<td>stat</td>
<td>100</td>
</tr>
<tr>
<td>762</td>
<td>pimpin</td>
<td>114</td>
<td>804</td>
<td>turn</td>
<td>106</td>
<td>846</td>
<td>electronica</td>
<td>100</td>
</tr>
<tr>
<td>763</td>
<td>peeze</td>
<td>114</td>
<td>805</td>
<td>sun</td>
<td>106</td>
<td>847</td>
<td>changes</td>
<td>100</td>
</tr>
<tr>
<td>764</td>
<td>honks</td>
<td>114</td>
<td>806</td>
<td>paper</td>
<td>106</td>
<td>848</td>
<td>business</td>
<td>100</td>
</tr>
<tr>
<td>765</td>
<td>hiphoppem</td>
<td>114</td>
<td>807</td>
<td>cuz</td>
<td>106</td>
<td>849</td>
<td>bandana</td>
<td>100</td>
</tr>
<tr>
<td>766</td>
<td>hero</td>
<td>114</td>
<td>808</td>
<td>care</td>
<td>106</td>
<td>850</td>
<td>already</td>
<td>100</td>
</tr>
<tr>
<td>767</td>
<td>ain</td>
<td>114</td>
<td>809</td>
<td>beatboxing</td>
<td>106</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>768</td>
<td>peacen</td>
<td>113</td>
<td>810</td>
<td>autotune</td>
<td>106</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
[Beginning with post #149658 from the MZEE forum]

A:
*Der Thread sucked*
*Achso: Moses rappt hin und wieder auf Englisch!*
*Oder hätte ich jetzt spricht hin und wieder rhythmisch. writen sollen*
*ist doch ein Bullshit thread!*
*Jeder wie er will, ist eben vieles sehr durch die V.S.von A. beeinflusst.*

A:
This thread sucks
Oh wow: Moses raps in English now and again!
Or should I have.. written [']speaks rhythmically now and again['].
this is a bullshit thread!
Just do what you want, a lot is influenced by the U.S. of A anyway.'

B:
-"Thread" ist eine amerikanische erfindung
-"sucked" zwar falsch geschrieben, weil die endung zum präteritum gehört, und eigentlich
versucht wurde, das wort in die deutsche grammatik inzubauen, aber trotz dummheit des
verfassers würde ich es als allgemein anerkannten anglicismus nennen; bleibt die frage welcher
ami das lutschen erfunden hat.
-"ist doch ein Bullshit thread!!";
bullshit ist ebenfalls ein anerkannter anglicismus, der schreiber hat versucht, es durch
großschreibung in die deutsche grammatik einzufügen, was sich beim zweiten teil seines
zusammengesetzten wortes nicht mehr fortsetzte. dieses zusammensetzen von begriffen ist rein
deutsch, nur eben wiedereinmal falsch durchgeführt, wie alles an diesem post.
also warum bezieht sich anglicismus nur auf gegenständlichkeiten, vA auf erfindungen? an dem
"bullshit-post" von gerade siehste doch nun, dass sogar deutsche wörter erfinden, die englisch
klingen sollen, nur um sich wichtig zu tun und zu zeigen, das sie mit beiden sprachen so ihre
probleme haben.

B:
-"Thread" is an american invention
-"sucked" is written incorrectly, because the ending belongs to the preterite, and while there's an
attempt to fit the word into german grammar, despite the idiocy of the author i would call it a
generally accepted anglicism; however the question remains, which american invented sucking.
-"ist doch ein Bullshit thread!"
bullshit is in likewise an accepted anglicism, the author tried, through capitalization, to fit it into
german grammar, an attempt which did not continue to the second part of his compound word.
this compounding of ideas is pure german, only again, falsely conducted, like everything on this
post.
so why does anglicization revolve around concreteness, above all on invention? on the "bullshit-post" here you now see, that even german words are invented that are meant to sound english just to seem important and to show, that there are problems with both languages.

A:
Was sollen die Beleidigungen, ich beherrsche Englisch durchaus sehr gut!
Es ging mir darum zu sagen, dass die Wörter jeder so nutzt in diesen Foren. Also die Diskussion lachhaft ist!
Dazu kommt - es heisst "sucked" oder eben "sucks" - wenn du es mit t schreibst, hast du es ja germanized.
Würdest du es "suckt" schreiben - weil mans so sagt?
Das sieht total komisch aus!

B:
was sind die Beleidigungen, ich beherrsche Englisch durchaus sehr gut!
Es ging mir darum zu sagen, dass die Wörter jeder so nutzt in diesen Foren. Also die Diskussion lachhaft ist!
Dazu kommt - es heisst "sucked" oder eben "sucks" - wenn du es mit t schreibst, hast du es ja germanized.
Würdest du es "suckt" schreiben - weil mans so sagt?
Das sieht total komisch aus!

A:
What are all the insults about, I speak English completely well!
I was trying to make the point that everyone uses the words this way in these forums. So the discussion is laughable!
For that reason - it's "sucked" or even "sucks" - when you write it with a t, you've germanized it.
Would you write it "suckt" - because you say it that way?
That looks totally strange!

B:
you're even using it like a german verb, so you've got to write it that way. just suckt. when you take a word from english and put it in a german sentence, then you have to germanize it, otherwise it's neither a german nor an english sentence. the declension is not part of the word, but
rather part of the grammar, and you can't bring that over. and a "Bullshit thread" is an example, because in english, there's absolutely no allowance for making those sorts of constructions. even if you didn't formally manage it, you've proceeded in a very german way.

this is all well and good, but i'm of the opinion that anglicisms are a good thing, when they add to the language, but when they just make garbage, it's a bad development. this wasn't meant as a personal attack. your post just happened to be nearby, but you will find such constructions in all kinds of threads. and the laziness in looking after one's own language is might be most visible in the genial quote from detlef soost (or "D!") [a German dancer]
"You guys are so tight, when I'm finished with you, it'll burn"
people like this should be stoned in the name of all the world's languages, because there's no way to come off dumber and more embarrassing.

A:
da sucken kein Dudenwort ist, kann ich es schreiben wie ich will
ich finde suckt sieht doof aus
drum sucked!
Was du davon hältst, ist mir egal.
Im Prinzip hätte man nur ein Wort hier hinschreiben müssen.
Als Antwort auf den Thread : "word" -- das ist typisch für dt. Foren.

A:
Since 'sucken' isn't in the dictionary, I can write it however I want
i think 'suckt' looks dumb
hence 'sucked'!
I don't care what you think about it.
In principle, one could have written a single word here.
To answer the thread: "word" -- that's typical for Ger. forums.
APPENDIX C: FINAL INTERVIEW PROTOCOL

[Read consent script, acknowledge agreement.]

1. Listening habits
   a. Wer sind deine lieblings-hip-hop-musiker (ggf. auch writer, breaker, producer, usw.)
      a. Who are your favorite hip hop musicians (or else graffiti artists, breakdancers, producers)
         i. Warum hast du X gern?
         i. Why do you like X?
         ii. Was sagst du zu hip hop musik, die dir gefällt?
         ii. What do you say about hip hop music that you like?
         iii. … die dir nicht gefällt?
         iii. … that you don’t like?

   b. Hörst du mehr amerikanische oder deutsche hip hop musik?
   b. Do you listen to more American or more German hip hop music?
      i. Prozentigerweise?
      i. Percentage-wise?
      ii. Wieso hörst du mehr amerikanische/deutsche hip hop musik?
      ii. Why do you listen to more American/German hip hop music?

2. Hip hop in Germany
   a. Wie bist du nach deine interesse in Hip hop (ggf. auch producing, breakdancing, writing) gekommen?
      a. How did you come to your interest in hip hop (or also producing, breakdancing, graffiti)?
   b. Was halten die meisten Leute hier in Hamburg, deine Meinung nach, von Hip hop?
      b. What do most people here in Hamburg think of hip hop, in your opinion?
         i. Ältere Leute/Jungere Leute?
         i. Older people/younger people?
         ii. Klassenkameraden/Familie?
         ii. Peers/Family?
         iii. Hast du mal Stress gekriegt, weil du dich für Hip hop interessierst?
         iii. Have you ever been given a hard time because of your interest in hip hop?
   c. Gibt es, deine Meinung nach, eine gewisse hip hop Sprache?
   c. Is there, in your opinion, a particular hip hop language?
      i. Wenn ja, was für eine Sprache ist das?
      i. If so, what kind of language is it?
      ii. Welche Wörter gehören dieser Sprache?
      ii. What words belong to this language?
APPENDIX C, CONT'D.

3. Attitudes about anglicisms
   a. Was hältst du von die Verwendung Anglizismen in hip hop Kultur hier in Deutschland?
   i. Kann es übertrieben sein?
      ii. Kannst du dich daran erinnern, wann es einmal übertrieben war?
         (freunde/musik)
   b. Was hältst du von die Verwendung Anglizismen im Alltagssprache oder Werbung (die vielleicht gar nichts mit Hip hop zu tun hat?)
   b. What do you think of the use of anglicisms in everyday speech or advertising (that maybe doesn’t have anything to do with hip hop?)
   b. Manche sprechen über ‚Deutschrap‘ oder ‚deutsche hip hop Kultur‘. Andere sagen, es gibt eher eine weltweite, internationale hip hop Kultur. Wie siehst du das?
   b. Some talk about 'German rap' or 'German hip hop culture'. Others say, there’s more like a worldwide, international hip hop culture. What do you think about that?

4. Prompts – asking for reactions to interview prompts
   a. Pyromaniac advertisement
   b. Franky Kubrick’s verse from '110-112':
      Es ist der big bad Boss, get back Dogg,
      Smoke dein Sticky weg, zip dein Six pack off, homie
      Ich spit mein Stuff, gib kein Fuck,
      Und push mir Heavyweights hit wie im Fitnessclub
   c. Excerpts from MZEE anglicism attitude thread (reproduced with translations in Chapter 7)

#1
„Ich krieg die Krise.....Ich seh MTV und dann sagen die: "Lasst die Crowd ausflippen mit eurem Sound, gewinnt 2 Tickets und rockt backstag." In dem Moment ist mir die Stellung von Anglizismen erst so richtig bewusst geworden. Seitdem acht ich drauf und es fällt immer mehr auf....Keine deutschen Wörter mehr sondern nur noch Anglizismen, Anglizismen, Anglizismen....Da isses kein Wunder, dass die deutschen Schüler zu blöd zum Scheißen sind....“
"Der Shit ist cool"

#2
„man kann sie hassn und wird trotzdem noch selber genug nutzen ohne es selber bewu ßt zu merken... man kann nix dagegen machen denk ich ma... leider, denn übertrieben nervt es wirklich! wobei gerade leude die sich mit hiphop und rap beschäftigen extrem gefährdet sind 😐"
#4
Ganz genau! Ich find wirklich unglaublich! Wirklich überall, "Lets go, zu großen Family Party mit vielen Boys und Girls in deinem Alter, finde neue Friends" ich hab mir schon mal ernsthaft darüber Gedanken gemacht, ob in 100 Jahren noch irgendjemand in Deutschland deutsch spricht... Obwohl wir ja laut einer These schon in 50 Jahren mehr Ausländer, als Deutsche in Deutschland haben, aber das is was anderes...
Warum machen wirs nicht so wie die Franzosen? Die haben das staatlich geregelt... schade, wenigstens die schlimmsten Aglizismen könnte man doch wohl öffentlich unterbinden. Wenigstens Anglizismen, für die es auch deutsche Wörter gibt, sollten "verboten" werden...
APPENDIX D: ORAL CONSENT PROTOCOL SCRIPT

“Hello. My name is __________ and I am a researcher in the Department of Linguistics at the University of Illinois at Urbana-Champaign. I am studying the use of English borrowings, that is, English words and phrases which are used in German. With your permission, I will be taking notes and recording [conversations/interviews]. The [conversations/interviews] will be roughly an hour in length. If you choose to participate in this research, neither your name nor any identifying details about you will be published or shared with others. Participation is voluntary, and you may choose not to answer any questions. You also have the choice at any time to have the audio recorder turned off, or to withdraw from the study.

Are you over 18 years of age?

Do you consent to participate in this research?

May I record your statements using an audiorecorder?”

German version:


Sind Sie über 18 Jahre alt?

Geben Sie Ihre Einwilligung, bei dieser Forschung mitzuwirken?

Darf ich Sie auf Band aufnehmen?”
University of Illinois
at Urbana-Champaign

Department of Linguistics

4080 Foreign Languages Building, MC-168
707 South Mathews Avenue
Urbana, IL 61801-3625

Marina Terkourafi
CC: Matthew Garley
Department of Linguistics
4080 FLB, MC-168

September 24, 2008

RE: A Study of Linguistic Borrowing in a German-language Hip-Hop forum on the Internet

Dear Marina,

This letter authorizes the use of human subjects in your project entitled *A Study of Linguistic Borrowing in a German-language Hip-Hop forum on the Internet*. The Department of Linguistics Institutional Review Board approved the protocol as described in your application. The expiration date for this protocol is **September 24, 2009**.

If you have not already done so, please provide the committee with hardcopies of your application on departmental letterhead along with proposal and any supporting materials. Please also remember to sign the application form. These documents will be kept in the departmental archive for future IRB audits.

Under applicable regulations, no changes to procedures involving human subjects may be made without prior IRB review and approval. The regulations also require that you promptly notify the IRB of any problems involving human subjects, including unanticipated side effects, adverse reactions, and any injuries or complications that arise during the project.

If you have any questions about the Linguistics Department IRB review process, or if you need assistance at any time, please feel free to contact any member of the Linguistics IRB Committee.

Sincerely,

[Signature]

Ryan Shosted
Member, Linguistics Department IRB Committee
February 18, 2010

Marina Terkourafi
Linguistics
4080 FLB

Dear Marina:

I am pleased to inform you that your project, “Linguistic borrowing in Internet forums,” was approved on February 15, 2010, for use of human subjects. You may now begin the data collection process.

Please be aware that this protocol was approved as is, any amendments to the protocol must be submitted for review. You are responsible for knowledge contained in the UIUC handbook, “Handbook for Investigators: For the Protection of Human Subjects in Research.” Any unusual events or problems with this protocol should be reported to the SLCL Human Subjects Review Committee.

This approval will expire on February 14, 2012.

Sincerely,

[Signature]

Andrea Golato
Associate Professor of German and Linguistics
Chair, SLCL Human Subjects Review Committee

xc: James Yoon
    Matt Garley
APPENDIX E, CONT'D.

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

May 5, 2010

Marina Terkourafi
Linguistics
4080 FLB
M/C 168

RE: Ethnographic research into borrowings from English in the Hamburg hip hop community
IRB Protocol Number: 10574

Dear Marina:

Thank you for submitting the completed IRB application form for your project entitled Ethnographic research into borrowings from English in the Hamburg hip hop community. Your project was assigned Institutional Review Board (IRB) Protocol Number 10574 and reviewed. It has been determined that the research activities described in this application meet the criteria for exemption at 45CFR46.101(b). Category 2 applies because the study involves interviews of adults in Germany who are involved in the hip hop community related to their use of English in German speech (called "English borrowings"). Consent of participants also allow the researcher to use tape conversational speech that takes place in public settings such as cafes (a form of observation of public behavior).

This determination of exemption only applies to the research study as submitted. Exempt protocols are approved for a maximum of three years. Please note that additional modifications to your project need to be submitted to the IRB for review and exemption determination or approval before the modifications are initiated. To submit modifications to your protocol, please complete the IRB Research Amendment Form (see http://irb.illinois.edu/cgi-bin/forms-and-instructions/research-amendments.html).

We appreciate your conscientious adherence to the requirements of human subject research. If you have any questions about the IRB process, or if you need assistance at any time, please feel free to contact me or the IRB Office, or visit our website at http://www.irb.illinois.edu.

Sincerely,

Sue Keuhn, Director, Institutional Review Board

c: Matthew Garley
APPENDIX F. TRACKLIST/DISCOGRAPHY FOR EPIGRAMS BY CHAPTER


