The Usage and Distribution of *so* in Spontaneous Berlin Kiezdeutsch

Stefanie Jannedy
Zentrum für Allgemeine Sprachwissenschaft, Berlin

In this paper I investigate the usage of the adverb and particle 'so' in spontaneous speech (interviews) collected from 21 speakers of the urban multi-ethnolectal youth language *Kiezdeutsch*. Speakers from the neighborhoods Kreuzberg and Wedding in Berlin are ranging in age from 14 to 18. The 1454 tokens of *so* available in the corpus (about 5 hours of speech) were classified into 10 different categories; some were structurally defined while others were defined along dimensions of meaning. Our current results indicate that there are differential usages patterns depending on the speaker's gender and age for some of these categories. Further, it appears that some patterns that have been attributed grammatical meaning may not appear frequently enough to establish a separate meaningful grammatical category. Rather, most instances of this kind of use of *so* appear to have a hedging function, indicating speakers' non-commitance to a specific circumstance.

1 Introduction

Labov's famous 1966 study on the social stratification of English in New York City marks the advent of urban sociolinguistics. As our world is becoming more and more global and connected in significant ways, migration and integration are challenges and chances at the same time as multi-cultural and multi-ethnical societies are emerging predominantly within urban areas. As multilingualism, cultural diversity and social integration are challenges to be mastered, we know that linguistic expression of individual style and group identity by young speakers in major urban areas are driving forces of linguistic innovation and language change which have lead to the emergence of new multi-ethnolects and distinct urban vernaculars. For example, Torgersen et al. (2006) showed that the locus of linguistic innovation and language change is inner-city East-London, an area with a large immigrant population. Even though urban areas in Germany are characterized by the multi-ethnicity of its population, differences in cultural heritage, and linguistic diversity, sociolinguistically informed quantitative investigations of multi-ethnolectal variation within and across different urban areas have not been undertaken in Germany.

In many European cities, researchers have noticed and studied the emergence of linguistic variation and the grammatical innovations introduced by young speakers from multi-ethnic urban neighborhoods (Multicultural London English: Torgersen et al., 2006; Kerswill et al., 2008; Straattaal (Netherlands): Nortier, 2001; Appel, 1999; Rinkeby-Svenska (Sweden): Kotsinas, 1992 and 1998; Bodén, in print; unpublished ms.; Kobenhavnsk multietnolekt (Denmark), Quist, 2005). However, most language research in Germany has largely neglected variation in speech based on sociolinguistic factors and predominantly focused on lexical or pronunciation variation along geographical and dialectal dimensions (e.g. Deutscher Sprachatlas: Herrgen, 2007).

Through waves of immigration during the 60s and 70s particularly from Turkey and Kurdistan (eastern Turkey), first generation speakers went through a more or less uncontrolled second language acquisition of German and learned a day-to-day variety of German simply by picking it up. Work on this so called *Gastarbeiterdeutsch* of the first generation immigrants was done (among others) by Keim (1978) and Pfaff (1981). Initially it was thought to be a variety or slang spoken by young people of Turkish decent only. Feridun Zaimoglu (1995) coined the in-group name *Kanak-Sprak* for the speech of adolescent males of Turkish descent which however appeared to have negative connotations for users outside of the group of speakers. Basic descriptive groundwork on the speech of second and third generation immigrants from Turkey and their mono-ethnic German peers has been laid by Androutsopoulos (2001) and Auer (2003) who used the terms *Türkendeutsch* and *Türkenslang* respectively to refer to this variety of German. Today there is a general agreement that an appropriate name for this variety of Germany needs to reflect the fact that speakers are young multiethnic urban speakers with a wide variety of language backgrounds (Auer & Dirim, 2004; Wiese, 2009; Krivokapic et al., 2010), so a term was coined that does not reference the speakers but the location where this variety is spoken: locally identified, tightly knitted neighborhoods all over Germany. The term *Kiezdeutsch* at best functions as a shortcut to invoking the notion of a highly stigmatized urban multi-ethnolectal youth language, often spoken in migrant communities in larger metropolitan areas in Germany which emerged on the basis of German and other languages such as Turkish and Arabic. For that matter, it is much more than an *inability* or *refusal* to speak (proper) German, it is more so an act of identity. It however completely neglects the fact that the term *Kiez* has a variety of different meanings, depending on location: in Berlin, a *Kiez* is a neutral term referencing a small local neighborhood, in Hamburg there is only one Kiez which happens to be the red-light district, and other urban centers such as Cologne or Munich do not have Kiezes at all. Thus, for the lack of a better term, the usage of *Kiezdeutsch* nowadays is common in the literature. *Kiezdeutsch* is not only characterized by phonetic or phonological alternations such as the realization of the palatal fricative /ç/ as post-alveolar fricative /ʃ/ (Auer, 2003; Mertins, 2010). While Turkish does not have palatal fricatives, we assumed that this sound is being substituted with a sound that is available in L1. However, our data also reveals alternations between the palatal and the post-alveolar fricative for some speakers, suggesting, that beyond substitution, other mechanisms for sound selection are at work. Other phonetic differences to German as spoken in Berlin or even more standard varieties are the avoidance of consonant clusters or differences in the realization of diphthongs. Auer (2003) and Wiese (to appear; 2009a) and others also describe various morpho-syntactic alternations, which we will not discuss here, the reader is advised to look at the references for literature on this issue and for examples of such alternations.

While collecting spontaneous lab-quality speech data through linguistic interviews from inner-city Berlin adolescents from Kreuzberg and Wedding, we noticed the pervasive use of the particle *so* 'like', occurring at the edges of phrases and phrase medially. It appears that *so* is being used in a wide variety of contexts and functions which will be explored in this study.

Example (1) shows instances of overuse of *so* by a 16-year old male German speaker of Turkish descent from Kreuzberg:

(1) Ich red mit dem Mann **so** ganz **so** locker spontan **so**
I speak with the man so very so cool spontaneously so
'I speak with the man like very cool and kind of spontaneously'

sehr **so** freundlich und **so**
very so friendly and so
'very like friendly and so on'

Another reason for classifying the usage and distribution of *so* is to model the duration of phrase-medial and phrase final *so* (Krivocapic et al., 2010) as to have means to correlate the duration data with the respective phrasal position. However, this work is discussed elsewhere. In this study, we set out to describe the usage-patterns, positions, and meanings of the particle *so* in this variety of German and will describe the distributions of the 1454 *so* in our database.

1.1 Meaning, Function & Classification of *so*

The particle *so* in German is multi-functional and the scope of meanings is difficult to capture which is evidenced by repeated attention it received during the last years (Umbach & Ebert, to appear; Hennig, 2007; Paul, 2008; Wiese, 2009a, 2009b). The Duden (2004) lists about nine different uses and meanings for *so*, among them as a deictic element, as an indicator of finality, degree or intensity, but also as a marker of a comparison or consequence. Hennig (2007) and Umbach & Ebert (to appear) address the problem of grammatical classification and part-of-speech membership of *so*. Hennig (2007) points out that the classification of *so* from reference works alone does not capture all meanings and that expressions containing *so* are hardly mentioned in theoretical works on this topic. She concludes based on her analysis of a random sample of roughly 50 tokens each of *so* from written text and spoken corpora, that a grammatical classification of *so* is rather difficult if not problematic because *so* often occurs in (idiomatic) expressions that are difficult to classify. She postulates the inclusion of the investigation of phonetic / intonational properties of *so* from spoken discourses to determine word-class membership and pragmatic meaning of this word. She notices that empirical work on such issues can point out problems and issues which would remain otherwise undetected in purely theoretical treatments of such a topic as we might not think of forms or usages of *so* that occur in spoken corpora of unscripted speech.

This however is a problem with many if not all empirically underpinned investigations (of part of speech classifications) from unscripted spontaneous speech: the corpus may not contain instances of all different kinds of use. We are well aware of these issues and by no means do we argue to have come up with an exhaustive list of occurrences and usage patterns of *so* in *Kiezdeutsch* in general. What we will show in section 2 of this paper is what seems to have emerged as sensible groupings from our corpus of spontaneous Berlin Kiezdeutsch. It is worth pointing out that *so* seems to be in some ways similar to the English *like* in that it can have grammatical function as an adverb but also discursive functions such as a discourse particle, a discourse marker or a quotative marker (Drager, 2010; D'Arcy, 2007).

Paul (2008), Wiese (2009a) and Wiese et al. (2009b) also recognize different usage of *so*: *so* can follow an argument as in *für Jugendliche so* 'like for adolescents', *mein Dings so* 'like my thing'; *so* can occur with prepositional phrases as in *so im Grünen* 'like out in the nature' or in *so aus Schöneberg* 'like from Schöneberg'; with adjective phrases *so blond so* 'like so blond' and it can occur with or precede an argument such as a bare noun as in *so Club* 'so club', *so Billardraum* 'so pool room', *so Naturtyp* 'so nature type'. Our ZAS-corpus of

Kiezdeutsch also shows such instances of use, thus we concur fully with what Wiese and colleagues state.

They further suggest though that some instances of use of *so* may serve to mark information structural prominence. In fact, they propose that in *Kiezdeutsch*, *so* is currently taking on a new and additional function, namely that of a focus marker". Wiese et al. (2009b:22) say about *so*:

> "[…] it can precede its argument […], follow it […], and it even occasionally brackets it […]. […], *so* in this usage is always combined with the focus constituent of the sentence, which carries the main accent. If one takes information-structural aspects into account, then, this seemingly erratic behavior can be subsumed under a unified account of *so* as a focus marker, a particle that attaches to the respective focus constituent in a sentence."

Thus, the authors attribute one of the functions or meanings of this particle to intensify the expression that is under the scope of *so* and lend it some kind of prosodic prominence. In fact, Wiese and colleagues have suggested such a function and claim the emerging or potentially grammaticalized function of *so* as a focus marker. They specifically mention the bracket construction whereby a *so* precedes and a second *so* immediately follows the argument ([*so* … *so*]).

Umbach & Ebert (to appear) provide a theoretical investigation of out-of-the-blue usage of *so* and argue that *so* is a demonstrative expression, combining with gradable and non-gradable expressions. They classify the usage of the German demonstrative *so* into three different groups: 1. deictic and anaphoric *so*; 2. intensifying *so*; and 3. hedging *so*. They suggest that *so* has an intensifying meaning that can be compared to *sehr* 'very' if it precedes gradable expressions such as adjectives as in (their example 3) *er ist so groß* 'he is so tall'. They further observe that *so* can combine with non-gradable expressions such as nouns (their example 4) *Ich möchte so Klammern* 'I want like clips'. In this usage, they propose, *so* expresses hedging and some kind of uncertainty about the appropriateness of the selected term. Consider the minimal-pair type example in which the *so* is unaccented and the last accent falls on the utterance final adjective *blau* as in 'Der Himmel ist so blau.' 'the sky is so blue' versus 'Das Kleid ist so blau.' 'the dress is like blue'. In the latter example, *so* is much more likely to receive a hedging interpretation.

Even though both groups of authors identify an intensifying meaning of *so*, they do not seem to agree on the meaning of *so* before non-gradable expressions. Thus, the interpretation of *so* + *noun* or any other type of argument (plus a following *so*) by Wiese and colleagues is in direct opposition to the interpretation proposed by Umbach & Ebert (to appear). Even though we have not classified occurrences of *so* according to gradable or non-gradable arguments, we have found and coded bracket constructions in our data. In this paper, we will offer a phonetic-phonological argument as to why we find the reasoning on the information-structural function and meaning of *so* offered by Wiese and colleagues not convincing.

The discussion in the existing literature was helpful for establishing our own classifications of *so*, however, it was still challenging to attribute meaning and function to all of the occurrences of *so* that we have found in our corpus. It is not our aim to add to the discussion on part of speech classification of *so*, we have merely looked to that body of literature to help us set criteria for our own classifications. These we deemed necessary to establish a level of description of the overall functions and meanings of *so* in this multi-ethnolect. In this study we will quantitatively investigate actual discourse usage patterns of *so* in a multi-ethnolect which is - among other features - characterized also by the over-use of this particle. The amount of data that we have collected from this multi-ethnolect allowed us to evaluate specific claims brought forward in the literature.

2 Methods

2.1 *Gminer*

To get a handle of the massive amount of spontaneous speech data, all recordings were first orthographically transcribed with a freeware audio-transcription tool *Transcriber* (version 1.5.1). The transcriptions are time-aligned with the audio-signal and anonymized programmatically. The transcription conventions such as the usage of punctuation (",", ".", "-" etc.) for different types of pauses were custom developed for this type of spontaneous data and adjusted on a need basis (Mertins, 2010). The output of *transcriber* plus the associated audio files were then uploaded into a browser based database search tool installed on a virtual server. This data mining tool is based on the *ONZE-Miner* (Fromont & Hay, 2008) which was originally developed to search through hundreds of hours of historical recordings of (the **O**rigins of) **N**ew **Z**ealand **E**nglish. The tool that we have used was localized for use with German data and we have named our data mining tool the *Gminer* (**G**erman miner).

The *Gminer* provides customizable search-spaces for adding speaker specific meta-information associated with that particular interview such as the age, gender, native languages, attended type/level of school, ethnicity, or neighborhood etc. of the speaker. Further, integrated into the *Gminer* is the German CELEX-dictionary, allowing for automatic canonical tagging of the lexical forms contained in the interview that was uploaded: automatically given are the phonological representation of each word form, the syntactic category, morphological structure, the overall word frequency and several other parameters available through CELEX. This meta-information can be displayed on separate layers in the transcription in the browser. The *Gminer* allows for sophisticated searches across words and across different layers whereby custom layers with specific annotations can also be added. A great advantage of this tool is the capability of downloading sound files associated with the search results for further annotation or segmentation in acoustic analyses software such as Praat. Krivokapic, Fuchs & Jannedy (2010) have used this functionality to first search, and then download hundreds of files and measure the duration of the /s/ and /o/ of the particle 'so' in phrasal-final positions to evaluate a data-driven analysis of different levels of the prosodic hierarchy.

Depending on the research question, search results (across several words and layers of annotation) as well as the associated meta-data (speaker information) can be exported into a spreadsheet and further marked up with relevant linguistic information (e.g. if the particle *so* is preceded by a noun or if it is following a noun etc.). This marked-up spreadsheet can be easily imported into *R*, a powerful statistics work package suitable for graphical and statistical exploration of large amounts of data.

2.2 Speakers

For the purpose of this study, we have extracted all instances of *so* from 21 speakers of Kiezdeutsch. 18 speakers were from Kreuzberg, only 3 from Wedding, thus, at this point we are not able to look for differences rooted in their local neighborhoods. As we have recorded 10 male and 11 female speakers, we are able to look for gender differences in the distribution of the data. Speakers were distributed across 5 different age groups ranging from 14 through 18 (1 x 14; 6 x 15; 8 x 16; 4 x 17; 2x18). The data was also coded also for factors such as *school form* attended, *native language* and *country of birth*. These factors however could also not considered at this point.

2.3 Categorization of *so* into ten different groups

In total, 1454 instances of usage of *so* have been extracted from the database. Each token was further tagged and annotated by hand for usage and function by the author and colleagues. We have abstained from theoretical assumptions of the use or grammatical group membership of *so* and tried to capture the actual meaning or the structural surroundings of this word. In accordance with syntax- and semantics experts and the existing literature, the following categorization criteria were established. It may be argued that in several ways these categorizations are oversimplifictions which gloss over more complex differences between the instances of use encountered in the data. Nevertheless have decided to used these criteria and categorization to make the crude point that usage patterns of *so* can either be structurally or semantically/pragmatically be defined. Further, it should be noted that we have not counted the same instance of use in different categories but made a choice what group to include this token with.

2.3.1 Categorization according to meaning

With this initial investigation, we have subsumed what Umbach & Ebert (to appear) call the hedging-*so* and the intensifying-*so* in a category that we have named *degree-modifier*. Instances of use in this group modify the degree of its argument. Examples for this category are *Türkisch kann ich auch nicht so gut* 'I can't speak Turkish so well' or *sind so viel Fragezeichen* 'there are so many question marks'. It is planned to further investigate these cases because naturally occurring language from spontaneous interviews may have forms that are not taken into consideration in theoretical deliberations on use and function. For example, there is an abundance of cases where the particle *so* appears after the argument and before a phrasal break, thus, clearly referencing the preceding material. However though, at this point we were not able to fully dissect this category into further subgroups.

All instances of *so* that occurred as reference to an object to which an entity was compared to were categorized as *comparison*. Examples are *so wie meine besten Freunde* 'just like my best friends', *ich fühle mich so als, als Berlinerin* 'I feel like a Berliner' or *so wie ein Deutscher* 'like a German'. Items were categorized as *correlate* when they related one state to another as in or *bei uns ist es so, dass* 'at our place it is like this' or as in *er will halt nicht so, dass ich Kopftuch trage* 'he does not want me to cover my head'. All cases of *so* that referenced something were categorized as *referential*. Examples of this category are *ich gucke sie immer so an* 'I look at her like that' or *war doch so!* 'it was like that!'. We were left with a category of miscellaneous items (*misc*) that were not readily classifiable. These occurred for example before pauses or in utterances that are characterized by false starts and such. Examples are *ja so ja äh.* 'yes, so yes uhm' or *so mh eigentlich so* 'so uhm, actually so'. We are well aware of the problem of potential ambiguities between categories and will have raters naïve to the purpose to this study as well as semantics experts reconfirm or dispute current judgments. We do expect however, that due to the relatively large sample size for spontaneous data, the overall distribution of categories in the data will remain relatively stable.

2.3.2 Categorization according to structure

Categorizations of *so* based on structure was somewhat less complicated than categorizations according to meaning. The five categories described here are mainly based on structural or co-occurrence descriptions, thus, they seem fairly straight forward with little room for dispute. We have classified all instances of *und so* where *so* occurs in immediate proximity following *und* 'and' as *coordinative* and all instances of *so* in immediate proximity following *oder* 'or' as *alternative*. The instances when *so* was directly followed by direct speech or quotations (orthographically marked in the transcribed data by adding a colon and quotation marks) were classified as *quotative* usages of *so*. Examples are *Ich so: "was macht ihr denn hier?"* 'I was like, what are you doing here?' or *da denkst Du so: "äh?!"* 'you're thinking like 'huh'?!'

Brackets are structurally defined through the occurrence of *so* before and after a word, sequence or argument as in the following examples: *die sind alle* ***so*** *verteilt* ***so*** *in der Türkei* 'they are all very dispursed like in Turkey'; *der geht* ***so*** *Berg hoch* ***so*** 'he like walks up a mountain'; *ist nicht* ***so*** *schwer* ***so*** 'is not so difficult like that' or ***so*** *Männergespräche* ***so*** 'like male talk'. The usage of *so* in structures like brackets supposedly combines with the focus constituent of the sentence (Wiese et al. 2009) and thus mark focus. Expressions such as *so oder so* 'so or so', *so und so* 'so and so' and *ach so* 'oh really' are subsumed in the category of 'bracket'. Note that the bracket construction of *so*___*so* was only counted as one instance of use of *so*.

2.4 Statistics

We conducted contingency table- and goodness-of-fit tests (chi-square analyses) with *age* and *gender* as independent factors and the number of counts produced for each of the ten categories of *so* as dependent variable. (The overall structure of our data generally allows for further analyses with factors such as *school-form*, *neighborhood*, *country-of-birth,* or *mother-tongue*. However, currently, some of the cells in the table were empty due to not having enough data and thus, no analyses were performed.

In those cases were the chi-square approximation calculation generated warning messages due to low counts in some cells, we ran several *monte carlo simulations*[1] with 10000 runs each. Each of these simulations generated different p-values, yet, the simulations consistently resulted in p-values that were reliably significant. Therefore, we can be sure that overall, the comparisons that involve

[1] Also see the R-help pages for chi-square analyses: *help(chisq-test)*. The actual R-command line is: *chisq.test(<table.name>.tab,simulate.p.value=TRUE,B=10000)*

cells with low counts are significant. Since no degree of freedom (df) is reported in these calculations, it can easily be identified where we ran the additional simulations. Further, in instances where we wanted to test for significant differences between factor levels (e.g. differences in usage of a particular *so*-usage-category by 16- vs. 17 year olds), we used a procedure, testing if the proportions are the same in different groups of data (R: prop.test).

3 Categorization Results & Usage Patterns

In the following section, we will show the distribution of the *so*-tokens into the 10 categories by showing raw counts in the graphs. Figure 1 shows a barplot for the overall distribution of the data into the categories. The categories are discussed in order of frequency of occurrence of the pattern in the data.

Most instances of use (548; 37.7%) of *so* are to modify the **degree** of its argument in cases such as *so schlimm* 'so bad' or *nicht so oft* 'not that often'. This use is also very well attested for standard varieties of German and seems to be the default use for *so* in German. The second largest group is comprised of instances where speakers have a **referential** use of *so* (330; 22.7%) in instances such as *bei uns ist so* 'with us that is the way it is' or in *so ein weiße Mütze* 'such a white cap'.

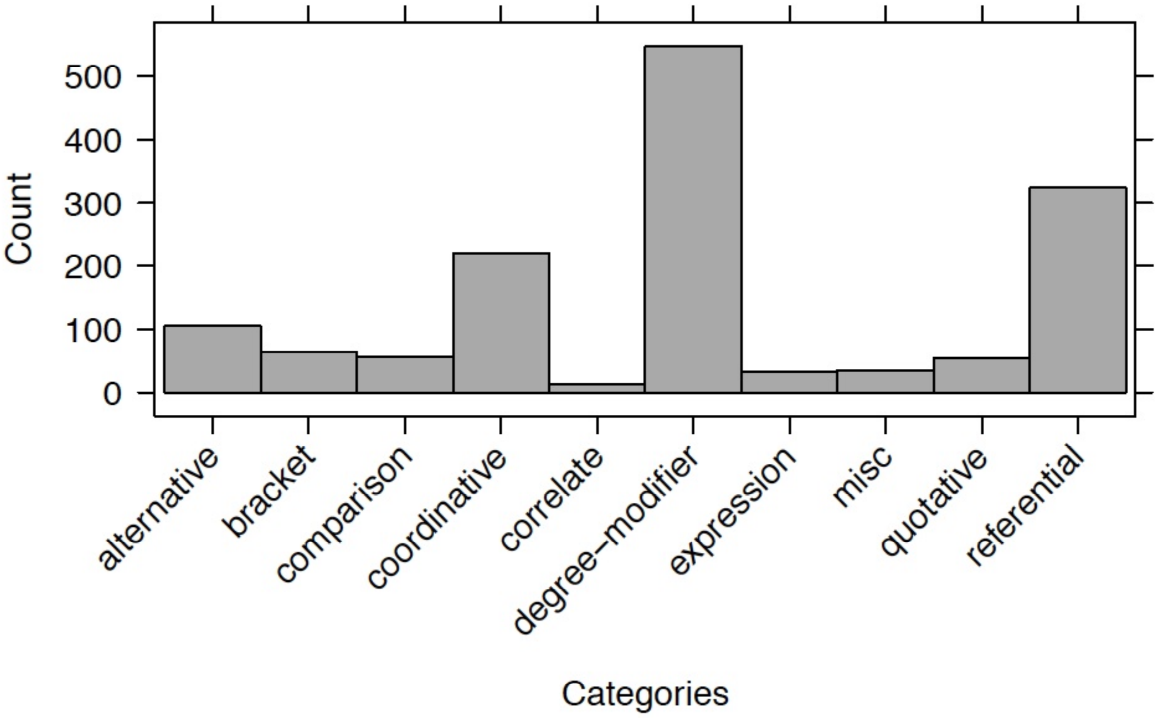

Figure 1: Cumulative graph of the differential use of *so* in the ZAS- Kiezdeutsch spontaneous speech database (raw numbers).

The use of *so* in a **coordination** *und so* ‘and so on’ occurred 222 times (15.3%), examples are *auf der Straße und so* ‘in the street and so on’ or *Schule, Universität und so* ‘school and university and so on’. Closely related is the category is the group we termed **alternative** *oder so* as it also combines a conjunction with the particle. We find 105 instances of use in the data (7.2%), examples are *dritter Monat oder so* ‘third month or so’ or *Türke oder so* ‘Turk or so’. To mark **comparisons** as in *wie so ein Tuschkasten* ‘like a paintbox’ or *Ach, Potsdam ist wie so ein Dorf ‘well, Potsdam is like a village’*, *so* was used 57 times (3.9%). The structurally defined **bracket** category occurred 66 times in the corpus (4.5%) – this category will be discussed in more detail below. **Quotative** constructions such as ... *und ich so:”Oh mein Gott”* ‘I was like: ‘Oh my God’ or *ich dachte so: “Nein!”* I thought like: ‘No!’ made up 3.7% (54 tokens) of the data in the corpus. The remaining three groupings are **correlates** (11 cases, 0.8%) like *so, dass* ‘so that’, **expressions** (32 cases, 2.2%) like *ach so* ‘I see’ or ‘ and a miscellaneous category that contained unclassifiable instances of *so* (29, 1.9%).

Dividing the data by gender reveals an overall effect with males generally using more instances of *so* than females (Pearsons χ^2 = 58.6765, df=NA, p<.001 with simulated p-value based on 10000 replicates).

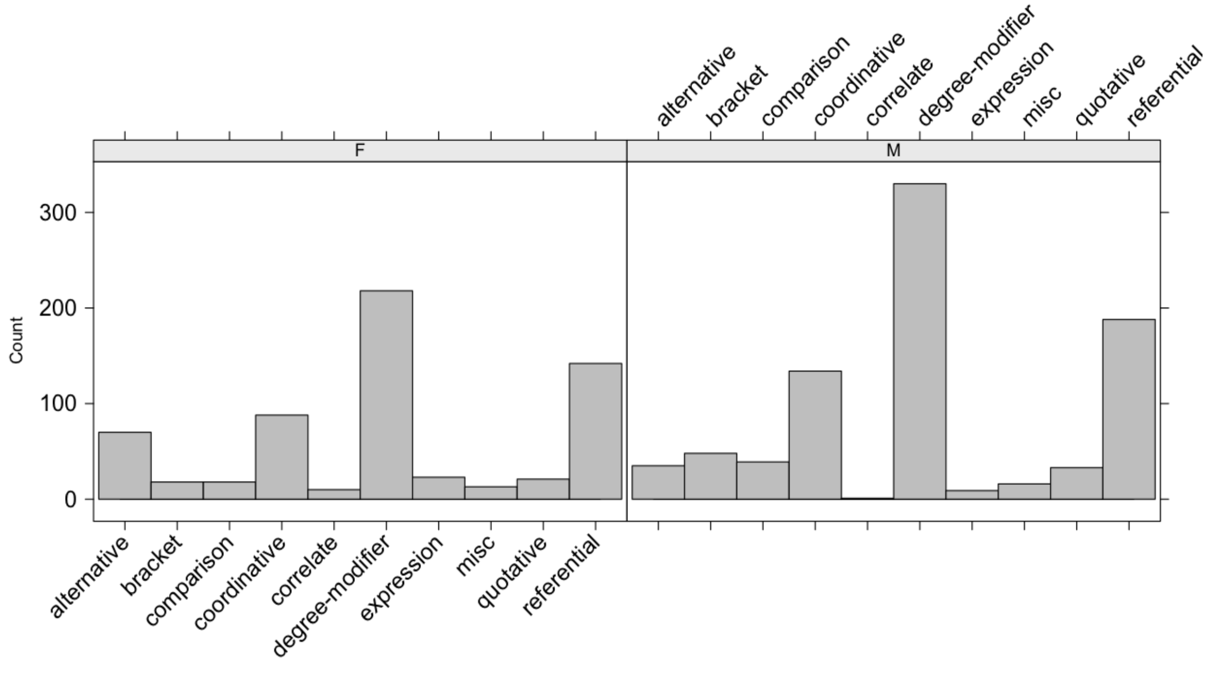

Figure 2: Cumulative graph of the differential use of *so* divided by the factor *gender* (raw numbers).

Individual comparisons for each category type by gender revealed significant effects for *alternative* ($\chi2$= 25.501, df = 1, p< .001) and *expression* ($\chi2$= 10.1885, df = 1, p< 0.002) with female speakers producing more tokens in these categories than males. The data also shows that males produce more *brackets* (χ2= 4.7806, df = 1, p=0.028) than females, suggesting that this structure may be an innovation predominantly used by males.

Comparing the data by age reveals an overall significant effect (χ2= 20.3349, df = NA, p< 0.001) with regard to the use of the bracket category. Data split up by category and age group is shown in Figure 3. This is not surprising given that we have already found a significant effect for gender and the group of 17 year olds is merely comprised of female speakers, under-using this linguistic innovation. A comparison of the proportions of usage of the bracket category by different age groups fails to reach significance between the 15- (13 instances of use) and 16- (44 instances of use) year olds (χ2= 3.3147, df = 1, p = 0.068). A comparison between the 16- and 17-year olds (4 female speakers, 0 instances of use of the bracket category), shows a significant difference (χ2-squared = 7.9202, df = 1, p< 0.01).

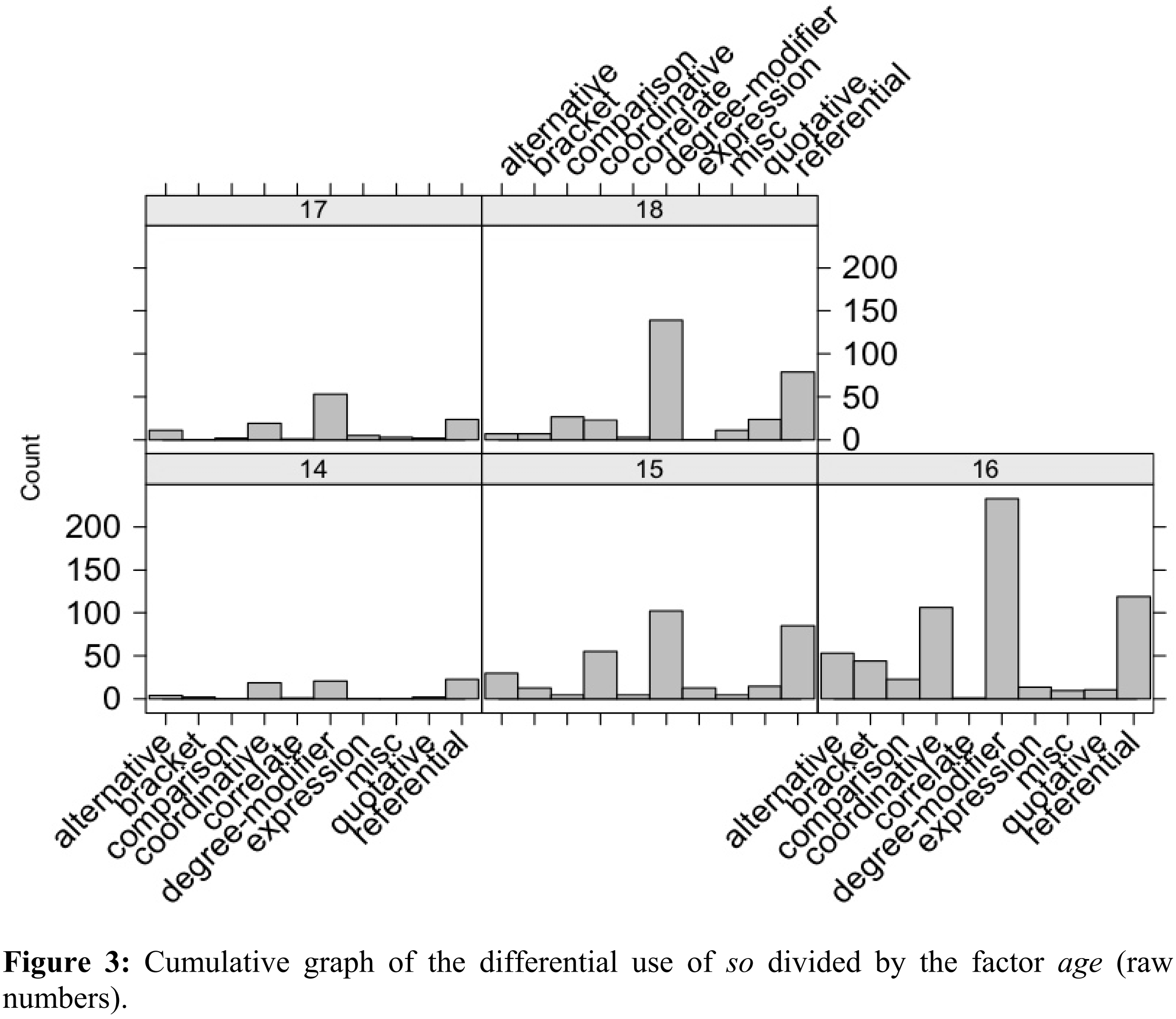

Figure 3: Cumulative graph of the differential use of *so* divided by the factor *age* (raw numbers).

The data suggests that usage of the bracket category (second bar from the left in each of the graphs in Figure 3) is gendered, it seems to be propagated especially by 15- (3M/3F) and 16- (5M/3F) year old males, whereas none of the 17 year old girls (0M/4F) produced such a token. The 18-year old male produced more bracket structures than the 18-year old female. Whether the data is truly age graded remains to be seen though, currently there is not enough robust evidence to make such claims given that at the time of analysis, there was only data from one 14-year old male and 2 18-year old speakers. As we are constantly adding data to our corpus, eventually we will be able to generalize over these age groups, too.

4 The ‘bracket’ Category

In total, there are 66 tokens in the ‘bracket’ category. Only 12 of the 21 speakers that we sampled for this study produced such a pattern. The structurally defined ‘bracket’ category itself is subdivided into three categories as shown in Figure 4: 1. coordinative structures (24) such as *so groß und so* ‘so tall and such’; 2. focus structures in which (according to Wiese, 2009; Wiese et al., 2009; and Paul, 2008) the particle *so* is proposed to serve as a focus marker (36) in this multi-ethnolect of German. An example is *Mit meinem Vater red ich eher über so Männergespräche so* ‘with my father, I speak about male topics if anything’ and further potential focus structures.

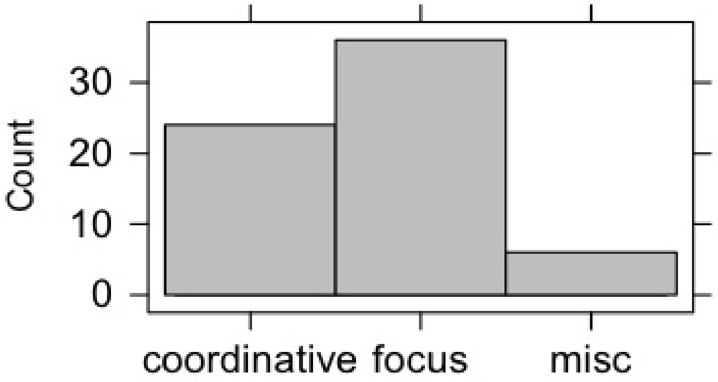

Figure 4: Cumulative graph of the differential use of *so* in the bracket category *so___so* in the Kiezdeutsch spontaneous speech database (raw numbers).

A third group includes all remaining structures that does not have great commonalities and is thus grouped in the miscellaneous category (6). Examples are expressions such as *so oder so* ‘so or so’ and *so und so* ‘so and so’ and other non-classifiable items such as *so weiter so machen* ‘continue to do so’. Figure 4 shows the entire set of bracket structures classified into these three sub-categories. A proportions test reveals that a single speaker produces significantly more focus structures than all other speakers together ($\chi 2 = 7.6768$, df = 1, $p < 0.01$). Thus, the extent to which we find *so___so* bracket structures that potentially fulfill the criteria for receiving a focus interpretation seem to be due to a speaker effect.

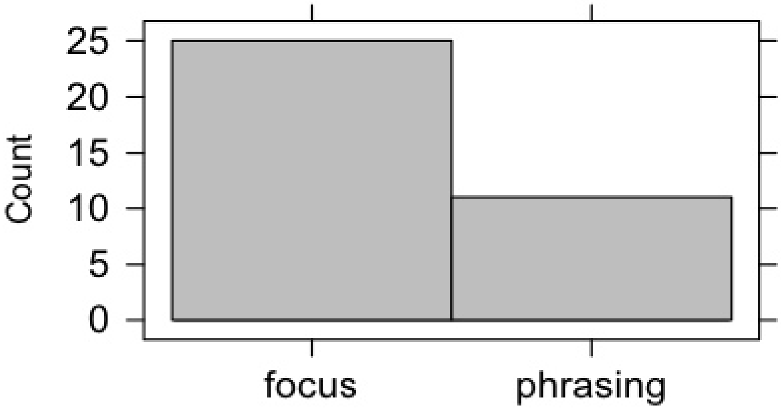

Figure 5: Cumulative graph of bracket category so___so in the Kiezdeutsch spontaneous speech database (raw numbers), divided by phonological phrase breaks.

A closer prosodic analysis of the 36 tokens that did satisfy the description of being a focus structure revealed that 11 of the utterances contained either an intermediate or an intonational phrase break either after the preceding *so* and the argument or between the argument and the following *so*, leaving 25 instances. Example (2) shown in Figure 6 shows a bracket structure which contains a phrase break after the initial *so* indicated by the lengthening:

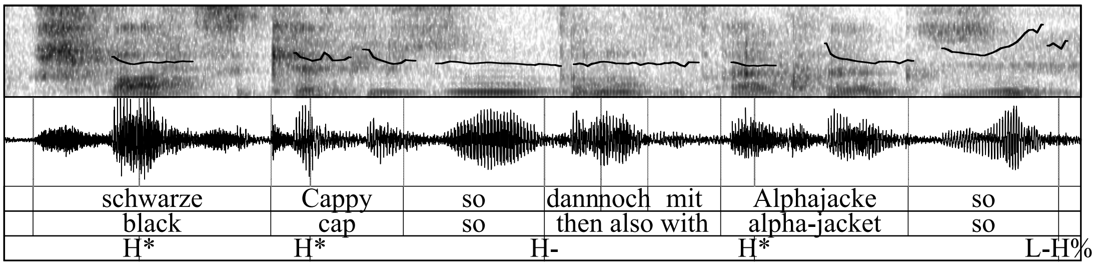

Figure 6: Spectrogram, waveform and f0-track for a bracket-sequence *so___so*, showing a phrase break as indicated by the lengthening of *so*.

(2) wenn man **so** // mit schwarze Cappy **so**
when one like - with black cap so
"[what impression does it leave] if you're with black cap

dann noch mit Alpha-jacke **so**
then also with Alpha-jacket so
and then also with an alpha-jacket on"

Another example that indicates phrasing includes *wenn man* ***so*** *arabisch oder türkisch //* ***so*** *laut spricht, dann haben die Angst [...]* 'when one speakers Arabic or Turkish rather loudly, they they are afraid'.

Some of the remaining 25 utterances are doubtful examples to the focus theory, too: *der geht so Berg hoch so* 'he walks like up the mountain' where the argument in focus would presumably be *Berg hoch* 'up the mountain' but where the accent in this case is located on *hoch* and *Berg* is unaccented. A second example is more or less untranslatable: *Und äh immer* ***so*** *Dings* ***so*** *halt* so 'and aeh, always something like' where the Argument in focus would be either the unspecified noun *Dings* or even the unspecific discourse particle *halt*. Given this non-specificity, it is unlikely that the speaker attempts to draw attention to the content of these lexical items. In ***so*** *erstmal* ***so*** *Ferien* 'so like first vacation', the focused element would be *erstmal*, a particle that also is very unlikely to have much attention drawn to it.

5 Summary & Discussion

While we have collected hours worth of data, at the present time, the database is not yet well balanced with regard to having comparable numbers of speakers for each age group, gender, neighborhood and native language. We recognize this as our shortcoming and ongoing and future work will remedy some of these issues: We are still recruiting speakers, interview them and add their data to our corpus for future analyses. While we are open to the criticism that our data does not adequately reflect all of the morpho-syntactic, lexical and phonetic/phonological variation that occurs in day-to-day interaction in the streets, we are confident that by now we have enough material that is of good enough quality that it lends itself to corpus analyses of spontaneous speech data and allows for generalizations over groups of speakers, especially in the domain of phonetics and phonology. For example, some of the data that we have collected is used for cross-dialectal perception studies (Jannedy, Weirich, Brunner & Mertins, 2010). Previous perception work on this topic was conducted with specifically created or recorded stimuli (see for example Niedzielski, 1999; Brunelle & Jannedy, 2010). Since our interviews also capture the interactional styles of the speakers, the data naturally lends itself also to investigations on the interface of morphosyntax and phonology which to a small degree we have exploited in this paper by evaluating the occurrences of *so* in bracket structures with regard to the phonological structure of the utterance as a whole.

We have found empirical evidence for the multiple uses, meanings and functions of the adverb and particle *so* in multi-ethnolectal Kiezdeutsch German. In about 500 minutes of spontaneous speech, there were 1454 occurrences of *so* (corresponding to an average almost 3 *so* per minute). Empirical evidence strongly suggests that there are gender, age and speaker effects in the usage of *so* whereby most usage patterns are well attested in standard varieties of German, too. One of these structures, the *so*___*so* bracket construction was proposed to function as a focus marker.

In examples like *Mit meinem Vater red ich eher über* **so** *Männergespräche* **so** 'with my father, I speak about male topics if anything' or in *Ähm, ich will* **so** *über Islam* **so** *[Vorträge geben]* 'I want to [give talks] about Islam', both taken from our corpus, it seems that an accent or at least some kind of acoustic prominence would fall on *Männergespräche* and also on *Islam*. This could be taken as evidence that within the bracket, there is accentual marking of focus. It is however the case that accents in languages like English and German often go on the last accentable constituent of an utterance. In a way, this is a default position for accent since it is much less marked than an accent early against a longer unaccented post-nuclear tail. It ought to be noted that not very accent marks a focus, and thus, in examples of the type given, there may be an accentual prominence which is unrelated to the pragmatic focus.

Moreover, we showed that a great proportion of the bracket structure was produced by a single speaker, calling into question the wide-spread distribution of this pattern or its rise to a grammaticalized pattern to indicate focus. All in all, of the 1454 occurrences of *so*, only 66 satisfied the structural description of 'brackets'. Within these 66 brackets, 36 satisfied the 'focus' structure (*so* .... *so*). And of these 36 that satisfied the focus structure, 25 had no phrase break (prosodic boundary) between the initial *so* and the argument and ultimately satisfied the structural description of these focus constructions. As some of the examples showed though, not all material enclosed in the *so* bracket is really meaningful. Further, even instances that structurally and prosodically fulfill the criteria may ultimately just do so because the default accent location is late in an utterance, thus, an accent on the argument enclosed in the *so* bracket that occurs late in an utterance may just receive a default accent rather than a focal accent. All in all, just about 25 of 1454 (1.7%) utterances contained the bracket-focus structure in this corpus. We call into question that this manifests a pattern, especially since the data shows that many of the bracket structures were produced by a single speaker.

There is a list of issues that have not been considered for the current scope of the paper. In the future though, we hope to address these: the categorization into *so* plus a gradable versus *so* plus a non-gradable expression; the phonetic-phonological categorization and implementation of *so* – when is it accented, when not, is the material following *so* always accented, is it only sometimes accented? If so, does it correlate with a specific structure or meaning? Based on the data that we have found and analyzed, we are not convinced of the emerging function of *so* as a focus marker in this multi-ethnolect. Rather, in most instances that do not have a clear meaning or function (quotative use or *so* before adjectives), we have associated a hedging interpretation with this particle, where the speaker refrains from being more specific about the argument and leaves much of the interpretation to the addressee of the discourse. This for example can also be tested in perception/rating tests with naturally collected data. Due to the pervasive use of *so* in Kiezdeutsch for some speakers, this multi-ethnolect lends itself well to an investigation of the durational properties of this particle in various prosodic positions within the utterances (Krivokapic, Fuchs & Jannedy, 2010). This work is in progress and will be discussed elsewhere.

Acknowledgements

I kindly acknowledge the diligent work and patience of my research assistant Micaela Mertins without whom none of the transcripts would have been created and tripple-checked. I thank my colleague Melanie Weirich, although much busy with her own PhD-research, who commented on this research at various times, and who was never too busy to discuss issues that arose with this work. Members of our lab made themselves kindly available for any help that was needed and their advice was and still is invaluable. The ZAS-Beirat and my colleagues spent time with me discussing this work, I thank them. Last but not least I thank our interviewees for their time, their stories and efforts - without them we would not have had any data to analyze. All errors remain my own.

References

Androutsopoulos, Jannis. (2001). *Ultra korregd Alder!* Zur medialen Stilisierung und Popularisierung von „Türkendeutsch". *Deutsche Sprache* 4/2001, 321-339.

Appel, René (1999). *Straattaal.* De mengtaal van jongeren in Amsterdam. Toegepaste Taalwetenschap in *Artikelen* 62;2,39-55.

Auer, Peter (2003). ‚Türkenslang'. Ein jugendsprachlicher Ethnolekt des Deutschen und seine Transformationen. In: Häcki-Buhofer, A. (ed.). *Spracherwerb und Lebensalter.* Tübingen. Francke, 255-264.

Auer, Peter & Dirim, Inci (2004). *Türkisch sprechen nicht nur die Türken: Über die Unschärfebeziehung zwischen Sprache und Ethnie in Deutschland*. De Gruyter.

Boersma, Paul & Weenink, David (2009). *Praat*: doing phonetics by computer (version 5.0.47) [Computer program]. http://www.praat.org.

Bodén, Petra (in print). Pronunciation in Swedish multiethnolect. In Svendsen, B. A. & Quist, P. (eds), *Linguistic Practices in Multiethnic Urban Scandinavia*. Multilingual Matters.

Bodén, Petra (unpublished manuscript). Adolescents' pronunciation in multilingual Malmö, Göteborg and Stockholm. In Lindberg, I. & Källström, R. (eds).

Boudahmane, Karim, Manta, Mathieu, Antoine, Fabien, Galliano, Sylvain & Barras, Claude (2008). *Transcriber*. A tool for segmenting, labeling and transcribing speech (version 1.5.1). [Computer program]. http://trans.sourceforge.net.

Brunelle, Marc & Jannedy, Stefanie (accepted). The Cross-Dialectal Perception of Vietnamese Tones: Indexicality and Adaptability. In Hole, Daniel and Elisabeth Löbel (eds.). *Linguistics of Vietnamese – an International Survey*. Berlin/New York: Mouton de Gruyter (TILSM series).

D'Arcy, Alexandra (2007). *Like* and language ideology: Disentangling fact from fiction. *American Speech* 82 (4), pp. 386–419.

Dirim, Inci & Auer, Peter (2004). *Türkisch sprechen nicht nur die Türken. Über die Unschärfebeziehung zwischen Sprache und Ethnie in Deutschland. Berlin*. De Gruyter.

Drager, Katie (2010). Sensitivity to Grammatical and Sociophonetic Variability in Perception. *Laboratory Phonology* 1(1), pp. 93-120.

Fromont, Robert and Hay, Jennifer (2008). ONZE Miner: the development of a browser-based research tool. *Corpora* 3(2), pp. 173-193.

Hennig, Mathilde (2007). So, und so, und so weiter. Vom Sinn und Unsinn der Wortklassifikation. *Zeitschrift für Germanistische Linguistik* 34, pp. 409-431.

Herrgen, Joachim (2007). From Dialect to Variation Space: The "Regionalsprache.de" Project. In The National Institute for Japanese Language: Geolinguistics around the World. *Proceedings of the 14th NIJL International Symposium*. Tokyo, pp. 75-80.

Jannedy, Stefanie, Weirich, Melanie, Brunner, Jana & Mertins, Micaela (2010). Perceptual Evidence for Allophonic Variation of the Palatal Fricative /ç/ in Berlin German. Poster to be presented at the Meeting of the *Acoustical Society of America* in Cancun, Mexico.

Kerswill, Paul, Torgersen, Eivind, Fox, Sue (2008). Reversing "drift": Innovation and diffusion in the London diphthong system. In *Language Variation and Change*, 20 (2008), Cambridge University Press, pp. 451-491.

Kotsinas, Ulla-Britt (1992). Immigrant adolescents´ Swedish in multicultural areas. In: Palgren, Cecilia, Lövgren, Karin & Bolin, Goran (eds.). *Ethnicity in Youth culture*. Stockholm. Stockholms Universitet, pp. 43-62.

Kotsinas, Ulla-Britt (1998). Language contact in Rinkeby, an immigrant suburb. In: Androutsopoulos, Jannis & Scholz, Arno (eds.). *Jugendsprache – lague des jeunes – youth language. Linguistische und soziolinguistische Perspektiven*. Frankfurt a.M. Peter Lang (=VarioLingua 7), pp. 125-148.

Keim, Inken (1978). *Gastarbeiterdeutsch - Untersuchungen zum sprachlichen Verhalten türkischer Gastarbeiter*. TBL Verlag Gunter Narr, Tübingen.

Krivokapic, Jelena, Fuchs, Susanne & Jannedy, Stefanie (2010). Prosodic Boundaries in German: Final Lengthening in Spontaneous Speech. Poster presented at the Meeting of the *Acoustical Society of America* in Baltimore, MD.

Mertins, Micaela (2010). *Phonetische Besonderheiten im Kiezdeutsch. Der palatale Frikativ*. MA-Thesis, Humboldt Univ. of Berlin.

Niedzielski, Nancy (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology* 18(1): 1-18.

Nortier, Jacomine (2001). „Fawaka, what´s up?" Language use among adolescents in Dutch mono-ethnic and etnically mixed groups. In Hvenekilde, Anne & Nortier, Jacomine (eds). Meeting at the Crossroads. Studies of Multilingualism and Multiculturalism in Oslo and Utrecht, 61-73. Oslo. Norvus Forlag.

Paul, Kerstin (2008). Grammatische Entwicklungen in multiethnischer Jugendsprache. Untersuchung zu Verwendung und Funktionen von *so* in Kiezdeutsch. Universität Potsdam.

Pfaff, Carol W. (1981). Incipient Creolization in Gastarbeiterdeutsch? An experimental sociolinguistic study. *Studies in Second Language Acquisition* 3:2.165-78.

Quist, Pia (2005). New speech varieties among immigrant youth in Copenhagen – a case study. In: Hinnenkamp, V. & Meng, K. (eds.) Sprachgrenzen überspringen. Sprachliche Hybridität und polykulturelles Selbstverständnis. Tübingen, Gunter Narr, 145-161.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, version 2.10.1. http://www.R-project.org.

Torgersen, Eivind, Kerswill, Paul, & Fox, Sue (2006). Ethnicity as a source of changes in the London vowel system. In F. Hinskens (ed.) *Language variation – European perspectives*. Selected papers from the third international conference on language variation in Europe (ICLaVE3), Amsterdam: Benjamins, pp. 249-263.

Umbach, Carla & Ebert, Cornelia (to appear). German demonstrative *so* – intensifying and hedging effects. *Sprache und Datenverarbeitung* (*International Journal for Language Data Processing*). (Draft Version 2009).

Wiese, Heike (to appear). The role of information structure in linguistic variation: Evidence from a German multiethnolect. To appear in Gregersen, Frans , Parrott, Jeffrey & Quist, Pia (eds.) *Proceedings ICLaVe5,* John Benjamins Publishing Company.

Wiese, Heike (2009a). Grammatical Innovation in multiethnic urban Europe: New Linguistic practices among adolescents. *Lingua* 119, 782-806.

Wiese, Heike, Freywald, Ulrike, Özçelik, Tina & Mayr, Katharina (2009b). Kiezdeutsch as a Test Case for the Interaction between Grammar and Information Structure. Interdisciplinary Studies on Information Structure (ISIS) 12. Universitätsverlag Potsdam.