Spatio-Temporal Prediction of Dialectal Variant Usage

Péter Jeszenszky
Department of Geography,
Ritsumeikan University
58, Komatsubara Kitamachi, Kita-ku
603-8341 Kyoto
pjeszenszky@gmail.com

Panote Siriaraya
Kyoto Institute of Technology
Matsugasaki, Sakyo-ku
606-8585 Kyoto
spanote@gmail.com

Philipp Stöckle
Austrian Centre for Digital Humanities,
Austrian Academy of Science
Postgasse 7-9
1010 Vienna
philipp.stoeckle@oeaw.ac.at

Adam Jatowt
Department of Social Informatics,
Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto
adam@dl.kuis.kyoto-u.ac.jp

Abstract

The distribution of most dialectal variants have not only spatial but also temporal patterns. Based on the 'apparent time hypothesis', much of dialect change is happening through younger speakers accepting innovations[1]. Thus, synchronic diversity can be interpreted diachronically. With the assumption of the 'contact effect', i.e. contact possibility (contact and isolation) between speaker communities being responsible for language change, and the apparent time hypothesis, we aim to predict the usage of dialectal variants. In this paper we model the contact possibility based on two of the most important factors in sociolinguistics to be affecting language change: age and distance. The first steps of the approach involve modeling contact possibility using a logistic predictor, taking the age of respondents into account. We test the *global*, and the *local* role of age for variation where the local level means spatial subsets around each survey site, chosen based on k nearest neighbors. The prediction approach is tested on Swiss German syntactic survey data, featuring multiple respondents from different age cohorts at survey sites. The results show the relative success of the logistic prediction approach and the limitations of the method, therefore further proposals are made to develop the methodology.

1 Motivation

Contact and isolation, in geographic space and in social space, are assumed to be the most important factors behind language change. The concept of *apparent time* (Bailey et al., 1991) hypothesizes that mother tongue is mostly acquired until the late teenage, after which one's language is more resistant to change. Throughout an individual's life contact patterns and social network might change (e.g., due to the ease of contact through media and changing migration or commuting patterns – especially from the 20th century). However, based on the apparent time hypothesis, if not uprooted, an individual's linguistic patterns can be assumed to reflect the contact patterns of their early life. With keeping all other variables constant, it can be assumed that for two people that are close in age and spent their youth near each other, the chance for a similar language is higher.

Thus, the quantification of contact possibility allows predicting current language usage and, through the concept of *'apparent time depth'*, future dialect change. If it is possible to predict the usage of variants based on the contact among users, core issues in sociolinguistics and diachronic linguistics such as the diffusion of variants, tracing back and forecasting change in language can be addressed with a better (spatial and temporal) granularity. Besides, through such an approach, linguistic theories long used, such as the apparent time concept (Bailey et al., 1991), language change following gravity-like paths (Trudgill, 1974) or wave-like diffusion (Yokoyama and Sanada, 2009; Blythe and Croft, 2012), can be tested. Further, it can contribute to natural language processing endeavours, such as predicting age from language attributes (Morgan-Lopez et al., 2017).

[1] An innovation is, of course, relative. A locally appearing new form with or without attestation elsewhere can be considered an innovation.

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 186–195
Florence, Italy, August 2, 2019.

This study, tracing language variation back to the patterns of contact between communities, contributes to existing approaches (e.g., Pickl and Rumpf, 2012; Wieling and Nerbonne, 2015; Yamauchi and Murawaki, 2016; Burridge, 2017) in language change and variation studies. So far linguistic geography mostly tested individual phenomena (Willis, 2017), but as obtaining data with better granularity becomes increasingly faster, computational approaches can speed up analysis in language change studies, and highlight variants that can be then more thoroughly investigated with the methods of qualitative and quantitative linguistics.

To account for the diverse roles of contact quantitatively, the relationship of the measured linguistic variation and variables affecting contact patterns – including social, demographic, policy-related or geographic factors – has to be tested. This paper is not the first step in this direction, with sociolinguistics and linguistic geography extensively having researched social status, geographic distances and trade, among others, in these regards (e.g., Labov, 1963; Gooskens, 2004; Nerbonne, 2009; Szmrecsanyi, 2012; Lameli et al., 2015). However, this paper shows one of the first steps towards assembling a model for predicting usage of dialectal variants, and thereby, language change by means of taking as many extralinguistic variables as possible into account. In this paper we start assembling the model by taking two main variable assumed by sociolinguistics to have a crucial impact on language contact and change: age and distance. In a previous paper (Jeszenszky et al., 2018), we provided first steps from the ordination aspect for assessing the spatial predictors of different grammatical domains.

The specific goal of this paper is to analyze the roles that age and distance play in language contact, as explanatory variables for the usage patterns of dialectal variants, tested at the linguistic level of syntax. We build a logistic predictor model at global and local scales for classifying multivariate syntactic data from a Swiss German dialect survey and present first results.

2 Materials and Methods

2.1 Dialect Data

It is often assumed in dialectology that of all linguistic levels, change in syntax is the slowest (Longobardi and Guardiano, 2009). It could mean that the association with age might be lower in syntax than for lexicon (Morgan-Lopez et al., 2017). However, the lower possible number of syntactic variants allows for more robust results with fewer responses in a survey.

The dialectal data used in this paper stems from the database of the *Syntactic Atlas of German-speaking Switzerland* (SADS; (Bucheli and Glaser, 2002; Glaser and Bart, 2015)). The database holds data collected in a series of four dialect surveys, which was conducted between 2000 and 2002, and probed 54 different (morpho)syntactic phenomena. At 383 survey sites, relatively homogeneously distributed throughout the German-speaking area, a total of 3'174 respondents (multiple respondents, 3-26 per survey site, median=7) filled in the questionnaires containing 118 questions. Respondents of several age groups (12-94 years old) were included at most survey sites. However, the age distribution is slightly skewed, with a median of 57 years (Stoeckle, 2018). The multitude of responses shows the local variation in variant usage, and give a higher attribute granularity and thus allows testing the association of variant usage and extralinguistic variables, such as age. Most survey questions involved translation from Standard German to the local dialect and multiple choice (MC) questions. For MC questions however, respondents could accept several answer variants as locally valid, and they were asked to specify their 'preferred' variants. In this work we rely on these preferred variants, as especially younger respondents tended to *accept* more variants (Glaser et al., 2019) – a clue for age as a factor conditioning usage patterns of dialectal variants. It has to be noted that even though dialectological research often refers to survey questions as variables, in this paper we call them *'phenomena'*, as the term 'variable' overlaps with the statistical terms used further on (i.e., explanatory variable, independent variable).

2.2 Predicting Dialectal Variant Usage Based on Age and Spatial Neighborhood

This paper presents the methodology and first results of our proposed approach for analysing the effects of age and regional contact. Regional contact is assumed to be more important in language change, manifesting itself in the variation of dialectal variants by age. We test the following two hypotheses:

- At the global scale, age explains the usage of dialect variants in linguistic phenomena.
- Age is a better predictor for the usage of dialect variants at the local scale.

Firstly, using logistic regression, similarly to Willis (2017), we analyse the predictive power of age at the global level, taking into account all respondents, for the usage of variants that correspond to dialectal phenomena. Secondly, we utilize a regionalisation approach: for every survey site s, taking a set of k nearest survey sites, we predict the usage of each variant in s, based on the age of respondents and the variant preference in the whole set.

Global scale. We test the association between linguistic variation as a categorical (nominal) variable and age as a continuous predictor variable, using logistic regressions. Logistic regression does not provide a good effect size statistic similar to R^2 used for Pearson's product-moment correlation. Nevertheless, its predictive power can be tested by training the logistic regression predictor on a training set in the data and checking whether the predictions of this model correspond to the observed data previously masked. We use a 10-fold cross-validation strategy, with all data used in the training set and all observed data predicted. This tests whether logistic regression based on age provides a significant prediction on the usage of variants at the global level, and if so, with what accuracy. Thus, we report in Figure 1 the significance in a binary way (i.e., whether the prediction of the usage of a certain variant is significant or not). Besides, we present the AUC in Figure 1 as well, as a typical performance measurement for binary classifiers, showing the separability, i.e., to what degree the model is capable of distinguishing between classes. The higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

Local scale. The regional approach can be viewed as a classification problem. Our model has to decide for each variant whether respondents at a central survey site s used it or not, based on age as the predictor variable in a set of k nearest neighbor survey sites. We use a logistic regression approach again. Using age as continuous and answer variants of all respondents as boolean variables, we train a logistic model and predict the variant usage for each respondent at s. We do this for all 383 survey sites. In this paper, we choose the k nearest neighbors based on Euclidean distance and we test models with different k values (5 to 50). Our approach employs distance cut-off, rather than distance decay, however it can also be assumed that the closer survey sites are, the more linguistic influence they have on each other.

3 Results

For this paper, we used 60 phenomena from the SADS survey (approximately half of all), which were already used in Jeszenszky et al. (2017). Appendix A provides some linguistic details on each phenomenon.

Results with regards to the explanatory value of age as a global predictor for variant usage are presented in Figure 1. For more than half of the variants considered, age is not a significant predictor (dark grey squares). The AUC values of separability, reported for the variants where the relationship with age is significant, are relatively low (0.5 means no discriminative capacity of the model). At the same time, variants that reach higher values typically have relatively few users (below 100 out of 3'714), e.g., *II5_3*[2]; 10 users and *II30_7*; 8 users. However, several variants with sparse usage are also found among those not predicted significantly by age. Variants with many users (e.g., *I12_1*; 2'683 users, *I7_3*; 2'880 users, *III2_1*; 2'021 users) typically have an AUC value between 0.5 and 0.6. These values of association between variant usage and age alone are relatively low overall, leading us to investigate the prediction power of age at the regional scale, the patterns of which are possibly concealed by the global patterns.

For each variant in each phenomenon, Figure 2 presents the number of survey sites (out of the total 383) in which age *significantly* predicts the variant's usage, based on $k = 13$ nearest neighbors. It is visible that age proves to be a significant predictor in a large amount or survey sites only for a few variants. These are, however, not always variants with a few users. The first few variants in each phenomenon usually cover the majority of respondents.

The distribution of one such variant (*III7_2*) is mapped in Figure 3 along with the significance and accuracy of the predictor variable age. The patterns in Figure 3 show that the higher number

[2]Variant coding includes the survey question number and a variant ID. For example, *II5_3* is Variant #3 in the 5th question of the 2nd survey sheet

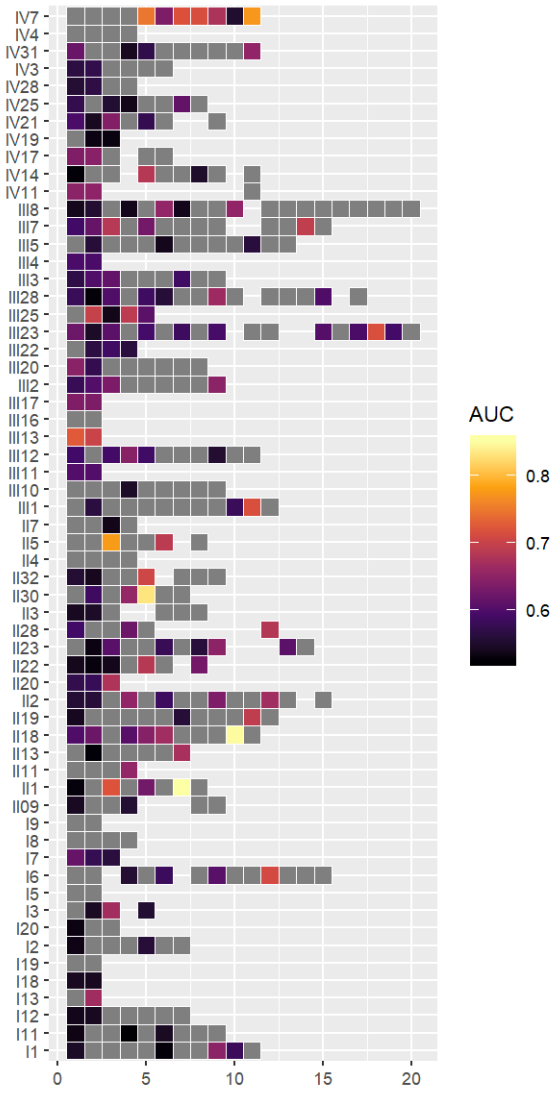

Figure 1: The global prediction power of logistic regression. The AUC values are plotted for each variant (horizontal axis) corresponding to the 60 linguistic phenomenon (vertical axis). Non-significance of the logistic regression is shown by dark grey squares.

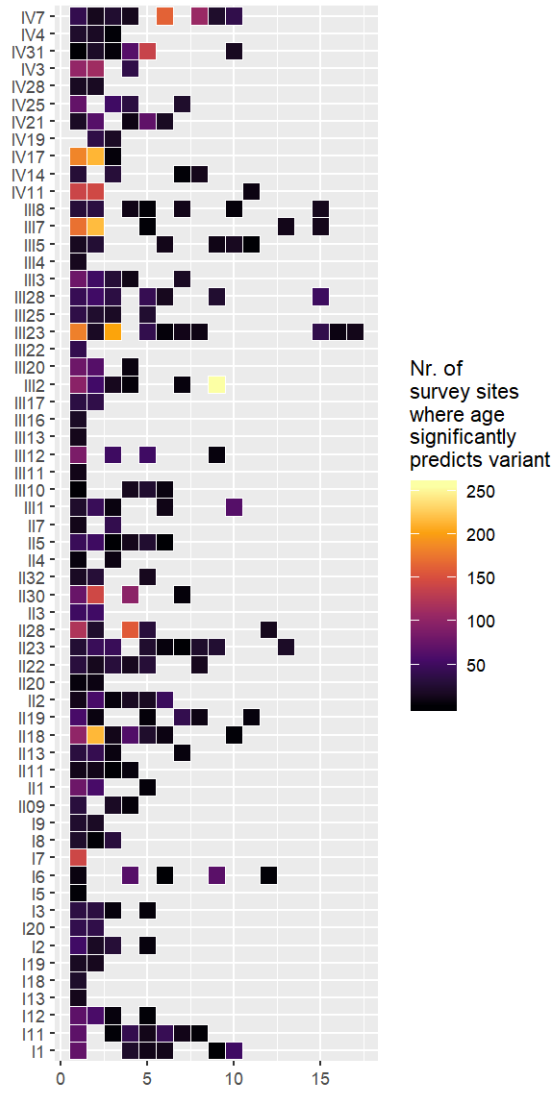

Figure 2: The local prediction power of logistic regression. For each phenomenon and variant, the colour corresponds to the number of survey sites for which the logistic regression on age proved significant, based on $k = 13$ nearest neighbors.

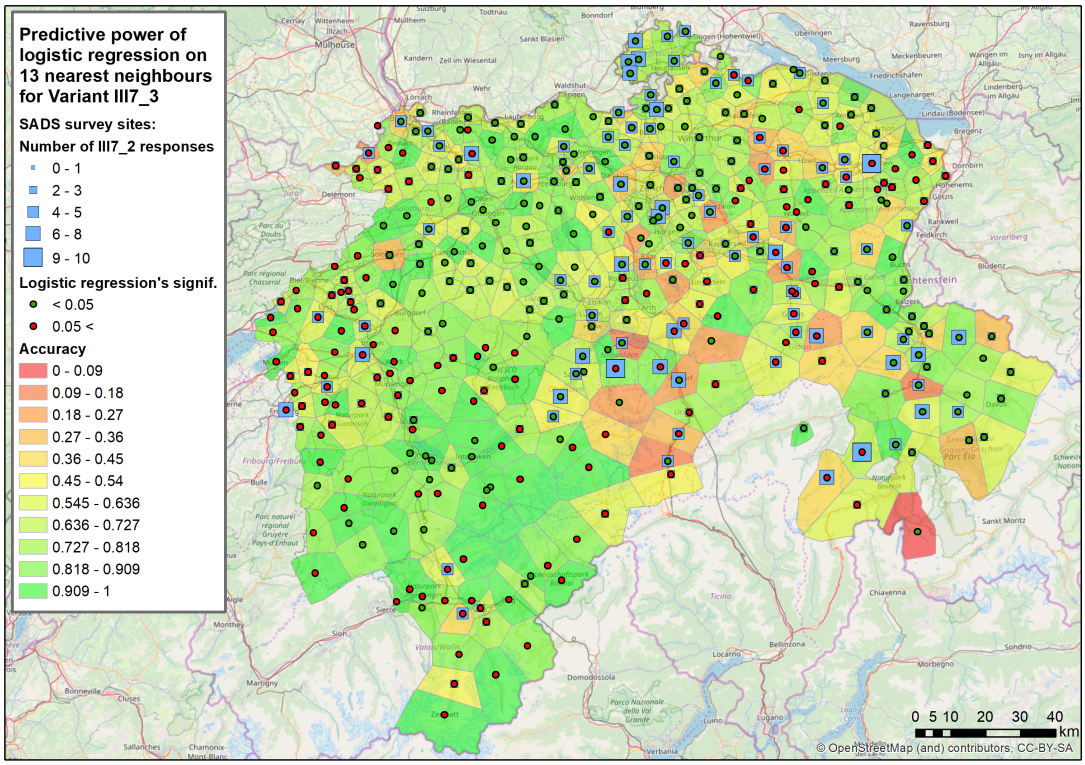

Figure 3: Mapping the significance ($p < 0.05$) and the accuracy of the logistic regression, for an answer variant *'hät s mer erzählt'* of Phenomenon *III7*, investigating the *'position of the personal pronouns'*, based on age and $k = 13$ nearest neighbors. Blue squares show the number of respondents using this variant. Accuracy is calculated by the proportion of correctly predicted usage.

of users does not necessarily make age a significant predictor. Significance of age as a predictor variable is spatially autocorrelated, which can be interpreted as follows. When present, the usage of this variant is characteristic of certain age groups at survey sites with green points, while at red ones it is used by different age groups.

As logistic regression is sensitive to class imbalances, it might not always be the best choice as a predictor when there are a lot of 0s and only a few 1s in the data, as it might result in false accuracy by predicting 0s only and not the 1s.

Interpretations of the first results show that age alone does not prove to be an exceptionally good predictor of syntactic variation. This is partly due to the nature of the data. It has been shown that while lexicon is more prone to have a correlation with age, syntax changes slower. The first results, however, show that already with a relatively simple approach, our research direction seems to be a worthwhile undertaking. Therefore, we have a wide outlook for further developing the methodology. The area and number of respondents involved in each model will be tested through different values of k, a distance decay approach and weights based on different parameters (including age). The spatial basis of the model will feature estimations of contact potential that have proved more 'informative' than Euclidean distance, such as travel time (Jeszenszky et al., 2017); linguistic gravity (Trudgill, 1974), predicting influence and therefore language change based on settlement populations as weights in a gravity equation; or linguistic distance (Pickl et al., 2014), assuming that the closer dialect varieties should be the outcomes of closer (historical) contact. Furthermore, different algorithms beyond the logistic predictor (e.g., random forests, SVM, XGBoost) will be tested in the prediction model.

Acknowledgments

We are grateful to Elvira Glaser, Gabi Bart and Sandro Bachmann of the Syntactic Atlas of German-speaking Switzerland (SADS) project for the provision and professional help with the linguistic data. Funding by the Swiss National Science Foundation (Project no. P2ZHP2_175019) is gratefully acknowledged. Further, we would like to acknowledge the comments of the anonymous reviewers.

References

Guy Bailey, Tom Wilke, Jan Tillery, and Lori Sand. 1991. The apparent time construct. *Language Variation and Change*, 3(1991):241–264.

Richard A. Blythe and William A. Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language*, 88(Number 2):269–304.

Claudia Bucheli and Elvira Glaser. 2002. The Syntactic Atlas of Swiss German dialects: Empirical and methodological problems. In Sjef Barbiers, Leonie Cornips, and Susanne van der Kleij, editors, *Syntactic Microvariation*, vol. 2. edition, pages 41–73. Meertens Institute Electronic Publications in Linguistics, Amsterdam.

James Burridge. 2017. Spatial evolution of human dialects. *Physical Review X*, 7(031008).

Elvira Glaser and Gabriela Bart. 2015. Dialektsyntax des Schweizerdeutschen. In Roland Kehrein, Alfred Lameli, and Stefan Rabanus, editors, *Regionale Variation des Deutschen. Projekte und Perspektiven.*, chapter 4, pages 79–105. De Gruyter, Berlin.

Elvira Glaser, Philipp Stoeckle, and Sandro Bachmann. 2019. Faktoren und Arten intrapersoneller Variation im Material des syntaktischen Atlas der deutschen Schweiz (SADS). In *Syntax aus Saarbrücker Sicht 3.: Beiträge der SaRDiS-Tagung zur Dialektsyntax 2018*, pages 1–30, Stuttgart. Steiner.

Charlotte Gooskens. 2004. Norwegian dialect distances geographically explained. In *Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe ICLAVE Vol. 2. 2004.*, pages 195–206, Uppsala.

Péter Jeszenszky, Sandro Bachmann, and Peter Ranacher. 2018. Towards the parameterisation and quantification of dialect contact potential: An extended abstract. In *GIScience 2018 unpublished extended abstract*, pages 1–6.

Péter Jeszenszky, Philipp Stoeckle, Elvira Glaser, and Robert Weibel. 2017. Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in Swiss German. *Journal of Linguistic Geography*, 5(2):86–108.

William Labov. 1963. The Social Motivation of a Sound Change. *¡i¿WORD¡/i¿*, 19(3):273–309.

Alfred Lameli, Volker Nitsch, Jens Südekum, and Nikolaus Wolf. 2015. Same same but different: Dialects and trade. *German Economic Review*, 16(3):290–306.

Giuseppe Longobardi and Cristina Guardiano. 2009. Evidence for syntax as a signal of historical relatedness. *Lingua*, 119(11):1679–1706.

Antonio A. Morgan-Lopez, Annice E. Kim, Robert F. Chew, and Paul Ruddle. 2017. Predicting age groups of Twitter users based on language and metadata features. *PLoS ONE*, 12(8):1–12.

John Nerbonne. 2009. Data-Driven Dialectology. *Language and Linguistics Compass*, 3(1):175–198.

Simon Pickl and Jonas Rumpf. 2012. Dialectometric concepts of space: Towards a variant-based dialectometry. In Sandra Hansen, Christian Schwarz, Philipp Stoeckle, and Tobias Streck, editors, *Dialectological and Folk Dialectological Concepts of Space - Current Methods and Perspectives in Sociolinguistic Research on Dialect Change*, linguae & edition, pages 199–214. Walter de Gruyter, Berlin/ New York.

Simon Pickl, Aaron Spettl, Simon Magnus Pröll, Stephan Elspaß, Werner König, and Volker Schmidt. 2014. Linguistic distances in dialectometric intensity estimation. *Journal of Linguistic Geography*, 2(01):25–40.

Philipp Stoeckle. 2018. Zur Syntax von afa (anfangen‘) im Schweizerdeutschen – Kookkurrenzen, Variation und Wandel. In *Syntax aus Saarbrücker Sicht 2. Beiträge der SaRDiS-Tagung zur Dialektsyntax*, pages 173–203, Stuttgart. Steiner.

Benedikt Szmrecsanyi. 2012. Geography is overrated. In Sandra Hansen, Christian Schwarz, Philipp Stoeckle, and Tobias Streck, editors, *Dialectological and Folk Dialectological Concepts of Space - Current Methods and Perspectives in Sociolinguistic Research on Dialect Change*, pages 215–231. De Gruyter, Berlin, Boston.

Peter Trudgill. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, 2:215–246.

Martijn Wieling and John Nerbonne. 2015. Advances in Dialectometry. *Annual Review of Linguistics*, 1(1):243 – 264.

David Willis. 2017. Investigating geospatial models of the diffusion of morphosyntactic innovations: The Welsh strong second-person singular pronoun chdi. *Journal of Linguistic Geography*, 5:41–66.

Kenji Yamauchi and Yugo Murawaki. 2016. Contrasting Vertical and Horizontal Transmission of Typological Features. *Proceedings of the 26th International Conference on Computational Linguistics (COLING-16)*, pages 836–846.

Shoichi Yokoyama and Haruko Sanada. 2009. Logistic regression model for predicting language change. In Reinhard Köhler, editor, *Studies in Quantitative Linguistics 5, Issues in Quantitative Linguistics*, pages 176–192. RAM-Verlag, Lüdenscheid (D).

A Appendices

A Appendix contains the 60 dialectal variables from the SADS in Table 1, 2 and 3, based on which the analysis was carried out.

SADS ID	Sentence *(Standard German)*	Sentence in English	Linguistic phenomenon
I.1	Entschuldigung, ich habe zu wenig Kleingeld, *um* ein Billett *zu* lösen.	Excuse me, I don't have enough change *in order to* buy a ticket.	infinitival purposive clause: linkage
I.2	*Wem* will er denn die schönen Blumen bringen?	*To whom* does he want to bring those beautiful flowers?	prepositional dative marking (PDM)
I.3	Oh, ich habe den Fritz *kommen hören*.	Oh, I *heard* Fritz *coming*.	perfect with 'hear': form and position of non-finite verb (IPP)
I.5	Der Korb ist *umgekippt*.	The basket *is toppled over*.	resultative: subject agreement
I.6	Wissen Sie, jetzt brauche ich sogar Tabletten *zum* einschlafen.	You know, now I even need pills *in order to* fall asleep.	infinitival purposive clause: linkage
I.7	Nein, das gehört *meiner* Schwester.	No, it belongs *to my* sister.	prepositional dative marking (PDM)
I.8	Aber ich habe im Fall schon gestern *geholfen abzuwaschen*.	But I already *helped doing the dishes* yesterday.	perfect with 'help': form and position of non-finite verb (IPP)
I.9	Also ich weiss auch nicht, ob er einmal *heiraten will*.	Well, I don't know if he ever *wants to get married*.	modal verb in subordinate clauses: position
I.11	Aber jetzt habe ich mich gerade hingesetzt, *um* ein Buch *zu* lesen.	But I just sat down *in order to* read a book.	infinitival purposive clause: linkage
I.12	Fischstäbchen muss man doch *gefroren anbraten*.	Actually, fish fingers should be fried while still frozen.	copredicative participle
I.13	Da *wird* gearbeitet.	*lit.* Here *will be* worked. (People are working here.)	expletive 'it' (impersonal passive)
I.18	Soll ich *welche* kaufen?	Should I buy some *of that*?	partitive object (pronoun)
I.19	Ich habe keine Ahnung, ob sie das Auto schon *bezahlt hat*.	I have no idea whether she *has* already *paid* for the car.	perfect auxiliary ('have') in subordinate clauses: position
I.20	Aber ich habe doch das Buch *dir* geschenkt.	But I gave the book as a present *to you*.	prepositional dative marking (PDM)
II.1	Hast du die Uhr *flicken lassen*?	Have you *had* the clock *fixed*?	infinitive particle (doubling/position) 'let'
II.2	Das ist doch die Frau, *der* ich schon lange das Buch bringen sollte.	This is the woman *to whom* I should have brought back the book long ago.	relative clause linkage: IO
II.3	Er *lässt* den Schreiner kommen.	*lit.* He *lets* the carpenter come. (He calls the carpenter.)	infinitive particle (doubling/position) 'let'
II.4	Du hast sicher viel *zu erzählen*!	You must have a lot *to tell*!	non-finite form with 'have to' (gerund)
II.5	Ihr dürft alles *liegen lassen*.	*lit.* You can *let* everything *lie*. (You can leave everything.)	infinitive particle (doubling/position) 'let'
II.7	Ich habe erst mit vierzig *fahren gelernt*.	I have only *learnt to drive* at forty.	perfect with 'learn': form and position of non-finite verb (IPP)

Table 1: The linguistic phenomena in SADS used in the experiments (part 1). The grammatical constructs of interest are highlighted in *italics*.

SADS ID	Sentence *(Standard German)*	Sentence in English	Linguistic phenomenon
II.9	Nein, sie ist gerade *verkauft worden*.	No, it *has* just *been sold*.	passive auxiliary and agreement
II.11	Er hat die Hand immer noch *eingebunden*.	He has his arm still *bandaged*.	resultative: object agreement
II.13	Du musst die Milch aber *heiss* trinken!	But you have to drink the milk *hot*!	copredicative adjective
II.18	Das ist der Mann, *dem* ich gestern den Weg gezeigt habe.	That's the man *to whom* I gave directions yesterday.	relative clause linkage: IO
II.19	Und dann ist ein Fuchs *geschlichen gekommen*!	And then a fox *came creeping* around!	verbal construction 'come' + motion verb
II.20	Ich möchte aber ein Auto, *das* ich auch bezahlen kann!	But I want a car *that* I can actually pay for!	relative clause linkage: DO
II.22	Nein, das ist *Peters* [Dreirad].	No, that's *Peter's*. [tricycle]	predicative possessive
II.23	Nein, das ist *Sandras* [Dreirad].	No, that's *Sandra's*. [tricycle]	predicative possessive
II.28	Das ist der Mann, *mit dem* ich immer schwätze.	That's the man *that* I always chat *with*.	relative clause linkage: PP
II.30	Der Hund *des Lehrers*	The *teacher's* dog	adnominal possessive
II.32	Ich habe Fritz *gesehen*	I have *seen* Fritz.	personal name: definite article and case inflection
III.1	Wenn es so warm bleibt, *fängt* das Eis *an* zu schmelzen!	If it stays this warm, the ice will *begin to* melt.	infinitive particle (position/doubling) 'begin'
III.2	*Wen* suchst du?	*Who* are you looking *for*?	interrogative pronoun: case
III.3	Für *wen* sind denn die Blumen?	*Who* are the flowers *for*?	interrogative pronoun: case
III.4	Die sind nicht für *dich*!	They are not *for you*!	personal pronoun (2sg): PP
III.5	Ich habe schon *angefangen* zu kochen.	I have already *started* cooking. (*lit.* have begun to cook)	infinitive particle (position/doubling) 'begin'
III.7	Sie hat *es mir* gestern erzählt.	She told *that* to *me* yesterday [about expecting a baby].	personal pronouns: position
III.8	Sie findet es nicht gut, dass ich *angefangen habe* zu rauchen.	She doesn't find it good that I *have started* smoking. (lit. *have begun* to smoke)	infinitive particle (position/doubling) 'begin'
III.10	Wenn sie dich erwischen, *bekommst* du den Fahrausweis entzogen!	If they catch you, you *get* your driver's license taken away.	'get'-passive
III.11	Also *mich* erwischt keiner!	Well, no one will catch *me*!	personal pronoun (1sg): DO

Table 2: The linguistic phenomena in SADS used in the experiments (part 2).

SADS ID	Sentence *(Standard German)*	Sentence in English	Linguistic phenomenon
III.12	Nimm die Suppe sofort weg, wenn sie zu kochen *anfängt*!	Take the soup off immediately, once it *begins* to boiling.	infinitive particle (position/doubling) 'begin'
III.13	Er gibt *sich* einfach keine Mühe.	He just doesn't put any effort into it. (*lit. for himself*)	reflexive pronoun (3sgm)
III.16	Die Strasse ist schon seit einem Jahr *aufgerissen*.	The street has already been *torn up* for a year.	resultative: subject agreement
III.17	Wir müssen *uns* das überlegen.	We have to think about it. (*lit. for ourselves*)	reflexive pronoun (1pl)
III.20	Er schaut nur für *sich selbst*.	He only thinks about *himself*.	reflexive pronoun (PP)
III.22	Sie ist grösser *als* ich.	She is taller *than* me.	comparative clause linkage
III.23	*Hinkend* ist er gelaufen.	He went home *limping*.	converb
III.25	Sie gehen halt lieber schwimmen *als* laufen.	They would rather go for a swim *than* for a walk.	comparative clause linkage
III.28	Dann ist er ja älter, *als* ich gemeint habe.	So he is older *than* I expected.	comparative clause linkage
IV.3	Ich habe *es ihm* schon geschickt.	I have already sent *it to him*.	personal pronouns: position
IV.4	*Wer* ist das gewesen?	*Who* was it?	interrogative pronoun: case
IV.7	Jetzt kannst du *anfangen*.	Now you can *begin*.	non-finite 'begin' with modal verb
IV.11	Doch, das ist im Fall *er* gewesen.	Yes, that must have been *him*!	personal pronoun (3sgm): subject
IV.14	Du musst das Licht anzünden, *um zu* lesen.	You have to turn the light on *in order to* read.	infinitival purposive clause: linkage
IV.17	Doch, das ist *er* sicher gewesen!	Yes, that was *him* for sure!	personal pronoun (3sgm): subject
IV.19	Ja, ich habe *etwas ganz* Schönes gekauft!	Yes, I have bought *something really* nice!	indefinite pronoun: position/doubling
IV.21	Ich habe nicht gewusst, dass er so spät fahren *gelernt hat*.	I didn't know that he *has learnt* to drive only so late.	perfect with 'learn': form and position of non-finite verb (IPP)
IV.25	Das glaubst du ja selber nicht, dass sie so früh lesen *gelernt hat*.	No way she *has learnt* to read so young!	perfect with 'learn': form and position of non-finite verb (IPP)
IV.28	Ich habe es (*dem*) Fritz gegeben.	I gave it *to* Fritz. (*lit. to the* Fritz)	personal name: definite article and case inflection
IV.31	Das *gefallen* täte mir auch!	approx. That would *do* to my *liking*! (I would like it, too!)	subjunctive auxiliary 'do' (position)

Table 3: The linguistic phenomena in SADS used in the experiments (part 3).