Examining the influence of pitch accents on word learning in German

Michael Walsh[1]*, Katrin Schweitzer*[1]*, Hinrich Schütze*[2]*, Dermot Lynott*[3]

[1]University of Stuttgart, Germany, [2]University of Munich, Germany, [3]Lancaster University, UK

michael.walsh@ims.uni-stuttgart.de

Abstract

This paper investigates the relationship between pitch and the lexicon in the context of a pitch-accented-word learning experiment in German. Participants were presented with novel abstract objects with nonsense words for names, and were required to remember these object-name pairs. The nonsense names were presented during training with either rising or falling pitch accents. In the testing phase participants were asked if auditory stimuli matched subsequently presented visual stimuli. In order to examine the effect of pitch accents on word learning, the auditory stimuli either matched or varied from their training equivalents with respect to pitch accent. The results show that this variation subtly influences reaction times despite the fact that German is not a tone language.

Index Terms: Lexicon; Pitch accents; Exemplar Theory; Word learning

1. Introduction

Traditional autosegmental-metrical models of intonation (mostly based on [1]) assume a clean separation between the lexical and the tonal level with regard to the assignment of intonation for Germanic languages. They assume that pitch accents are specified predominantly according to top-down information such as syntactic or semantic factors. That is, categories that are assumed to be phonologically different are assigned to a sequence of words taking the metrical structure of the utterance [2] into account. Phonetic implementation rules then determine the exact shape of the pitch contour. The word level is assumed to have very little influence on the realisation of an accent, apart from micro-prosodic effects (e.g. [3]).

This perspective is well accepted in the research community. However, it is at odds with a branch of research demonstrating that speakers of a language store acoustic detail of individual instances of previously perceived stretches of speech in memory, and employ it for production and perception. In production, the stored instances serve as production targets, while in perception/categorisation these stored exemplars act as references to which new stimuli are compared. These usage-based, or exemplar-theoretic accounts of speech processing [4–7] assume a much tighter cohesion between the word level and the tonal level: fundamental frequency, perceived as pitch, is part of the spectral information of any perceived instance (e.g. [8, 9]). Additional evidence in keeping with an exemplar-theoretic perspective can be found in psycholinguistic [10, 11] and machine-learing pitch accent prediction studies (e.g. [12, 13]).

Building on the evidence outlined above, this paper investigates the possibility that the word and tonal levels are, at least to some extent, coupled.

2. Experiment

The following experiment sought to determine, via a word learning task, if tonal contours can form part of lexical memory. The experiment involved two phases, training and testing, which all participants underwent. Participants were not in any way aware of the purpose of the experiment.

Materials Pictures of six abstract objects were employed for the study [14]. Each of these objects was assigned a nonsense word composed of two CVCC syllables of German. The use of nonsense words ensures that no exemplars of the words exist in participants' mental lexica. However, given that syllable frequency is known to influence lexical retrieval (e.g. [15]), syllable frequency (assigned based on [16]) was controlled for in the stimuli. Each word was then embedded in a carrier sentence in sentence-medial position and the sentence was recorded once with a rising, and once with a falling, accent on the nonsense word, which was produced with a trochaic stress pattern. The German carrier sentence conveyed the following meaning: "Please observe the *X* on the screen" where *X* represents the nonsense word. Thus, for example, the sentence "Bitte betrachten Sie den *Dohltbuept* auf dem Bildschirm" was recorded twice, the first time with a rising pitch accent (L*H in the GToBI(S) notation [17]), on Dohltbuept and the second time with a falling accent, H*L, on Dohltbuept. To ensure that both L*H and H*L sounded natural in this context, the sentence-final word *Bildschirm* (screen) always carried a falling accent. The recordings took place in an anechoic chamber and were produced by a female speaker of standard German, an expert phonetician, experienced in producing a target intonation contour. All sentences were examined for tonal accuracy, consistency, naturalness and, in particular, position of word stress (syllable-initial), by trained phoneticians. The resulting dataset for the experiment comprises 12 sentences, 1 for each pitch accent condition (rising and falling) for each of the 6 nonsense word-object pairs. The stimuli for the training and testing phase were con-

Table 1: *Training materials: Nonsense words, their training pitch accents and the frequency of their syllables*

Name	Pitch Accent	Syllable Freq.	SAMPA
Dohltbuept	L*H	Low	"do:ltbYpt
Murpbaehnf	L*H	Low	"mURpbE:nf
Wahltfantz	L*H	High	"va:ltfants
Kuenfwackt	H*L	High	"kYnfvakt
Zenntwaehnt	H*L	High	"tsEntve:nt
Poltzhohmt	H*L	Low	"pOltsho:mt

structed from these recordings. The experiment was carried out using Slide Generator software [18].

Participants Forty participants (20m, 20f) took part in the experiment. They were all native speakers of German with no hearing/sight impairment.

Procedure For the training phase, participants seated in front of a computer screen and wearing headphones (AKG, model HSC 271) read instructions informing them that they would be presented with a number of novel objects, one after the other, and that they should try and remember the name of each object. Each participant first heard a carrier sentence in which the novel object's name was embedded, immediately followed by the appearance of the object on the screen. Training stimuli were presented in 5 blocks. Each block comprised six sentences, with each sentence containing one of the 6 word-object pairs, in a randomised order. That is, in total, participants heard 30 sentences, and were exposed to each of the 6 word-object pairings 5 times. Importantly, participants only ever heard a word-object pair with a rising accent or a falling accent, but not both conditions. Specifically, half of the nonsense words were always presented with a rising accent in the training phase, and the other half were always presented with a falling accent (see Table 1).

In the testing phase, participants were presented with on-screen instructions stating that in the next phase they would hear the name of one of the objects they had learned and would be presented with an object on screen. If the name and the object matched they should press the "yes" key (right arrow key), otherwise they should press the "no" key (left arrow key). Participants were told to respond as quickly as possible. If the participants did not respond within 3 seconds of a particular stimulus, a time-out value was recorded and the next stimulus was presented.

In this phase participants heard word-object pair stimuli with both rising and falling accents, to determine if hearing a nonsense word with an accent different to that perceived during training would affect the participants reaction times. The presentation of stimuli was randomised for each participant and a total of 120 trial stimuli were employed, 20 per nonsense object, 10 of which were instances where the object was pronounced with a rising accent and 10 of which were it pronounced with a falling accent. For each object-word pair, in half of the presented instances the audio and visual stimuli matched.

Consequently, for a given object image, the stimuli for the test phase vary under four different conditions with respect to how well they match the respective stimulus from the training phase (see figure 1). Each stimulus in the test phase could either match or mismatch the stimulus from the training phase on the *nominal level* (either the object name was the same or different from the one learned in training) or match or mismatch it on the *tonal level* (the pitch accent on the object name was either the same or different from the one heard in training). Each of these 4 conditions occurs 5 times per nonsense object. For instance, for the picture denoting a *Dohltbuept*, the stimulus in the testing phase could either be exactly the same (subjects hear the word "Dohltbuept" realised with the same accent as in the training phase (L*H)), or the stimulus could be a partial match where the word is the same ("Dohltbuept"), but the accent is not (H*L), or the stimulus could be a partial match, where the word is not the same (any other object name, e.g. "Kuenfwackt") but the accent matches (L*H), or it could be a total mismatch, where neither word nor accent match the conditions from the training phase.

Identical stimuli were not contiguously presented to prevent learning during this testing phase. On each trial the participants response and reaction time was recorded.

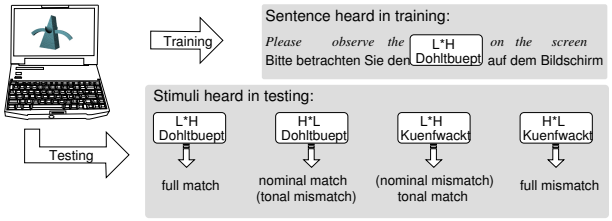

Figure 1: *Experimental conditions for example* Dohltbuept

Data The obtained dataset comprised 4800 observations (120 tests from each of the 40 participants). Of these, in 3758 cases, the subjects had judged correctly within the 3 second window (the remainder of the data comprised 267 timeouts and 775 misjudgements).From this dataset, outliers, defined as reaction time values that fell outside the whiskers in a boxplot, hence that were more than 1.5 interquartile range away from the quartiles, were removed. This reduced the dataset to 3624 observations, that is, 4% of the observations were considered outliers.

3. Results

Our analysis examined participant reaction times where the response was correct, i.e. the participant correctly judged that an object-word pair was a match or a mismatch. Error rate analysis yielded no significant effects with regard to the tonal level.

To determine the effect of the experimental conditions on log response latency, a mixed-effects linear regression model was fitted to the data with participants and objects as random effects. Match or mismatch on the nominal level, match or mismatch on the tonal level, number of the trial, and frequency of the syllables were incorporated as fixed effects, as was the interaction between the nominal and the tonal level[1].

To achieve a normal distribution, the reaction times were transformed to their logs. We checked for normality and homogeneity by visual inspections of plots of residuals against fitted values. To assess the validity of the mixed effects analyses, we performed a likelihood ratio test comparing the model with fixed effects to the null model with only the random effects. The model including fixed effects differs significantly from the null model at $\alpha = 0.001$.

There were significant effects of nominal (mis)match, trial and syllable frequency, and we also observed a significant interaction effect between nominal and tonal factors. That is, these fixed effects demonstrated a significant influence on log latencies at a sigificance level of $\alpha = 0.05$ or lower for the p-values [19]. Table 2 gives an overview of estimated coefficients, standard errors and t-values of the significant fixed effects. We detail patterns of these effects below.

Effect of tonal realisation The effect that the tonal realisation has on log latencies becomes evident when looking at the interplay between the object name and the accent realised on the object. Figure 2 illustrates the interaction between the tonal

[1]Incorporating interactions with trial and the other fixed effects did not significantly improve the model.

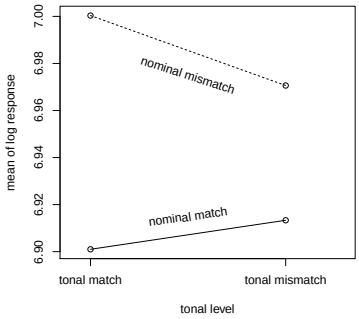

Figure 2: *Interaction plot showing the relationship between nominal match/mismatch and tonal match/mismatch conditions with respect to log reaction times.*

level and the word level. The figure displays the mean values of the logged responses for the four conditions: full match 6.901 (1067ms), nominal match but tonal mismatch 6.913 (1084ms), nominal mismatch but tonal match 7.0 (1167ms), full mismatch 6.970 (1137ms). The upper line shows those cases where the objects and the word did not match. In those cases, subjects were generally slower than in cases where the word denoted the correct object (lower line). The slope of the lines illustrates the effect that mismatch on the *tonal* level has on response times: participants were faster at correctly judging that a word-object pair matched if the pitch accent contour which they heard in training was also present (lower line). Furthermore, the figure also highlights the fact that participants were faster at correctly rejecting word-object pairs that did not match if the pitch accent contour on the presented word did not match that presented in training (upper line). In order to further investigate the potential impact of tonal information two Welch Two Sample t-tests (one-tailed planned comparisons) were performed. The first test compared the means of the tonal match and mismatch groups in the case of nominal matches (figure 2 bottom line), yielding no significant result (t=-0.678, p=0.248). The second test compared the means of the same groups but this time in the case of nominal mismatches (figure 2 top line) yielding a significant effect (t=1.792, p=0.036). These results are discussed below.

Effect of syllable frequency Syllable frequency had a significant effect on logged reaction times: subjects reacted faster when the object's name consisted of frequent syllables than in the infrequent case (mean of logged response for stimuli comprised of frequent syllables was 6.933, as opposed to 6.963 for the infrequent case).

4. Discussion

Our results show that subjects are significantly faster at correctly rejecting mismatching word-object pairs when the pitch accent presented in the testing phase deviates from that presented in training. For example, if in training participants heard *Dohltbuept* with a rising accent and saw a given object, they were significantly faster at rejecting *Dohltbuept* as the name for another object, when the accent on *Dohltbuept* was a falling one. However, as outlined above, we anticipated a delay when the object was presented with the correct name, but with a pitch accent not heard in training, i.e. the nominal match but tonal mismatch condition. A delay in reaction time is present in figure 2 (difference between the bottom left and bottom right), however it is not a significant delay. Thus, our results show that the tonal pattern acquired during learning appears to have some impact on word processing time, albeit a limited one.

	Estimate	Std. Error	t value	
nominal (mis)match	-0.1272457	0.0150330	-8.46	***
trial	-0.0023372	0.0001539	-15.18	***
syllable frequency	0.0340716	0.0113564	3.00	**
interaction tonal*nominal	0.0416285	0.0209822	1.98	*

Table 2: *Results of the linear mixed model predicting logged responses; signficant fixed effects.* *** $p < 0.001$*;* ** $p < 0.01$*;* * $p < 0.05$

The findings, therefore, point to a potential coupling of lexical and tonal information in how novel objects are remembered. This is unlikely to be due to pragmatic considerations as the object names are presented in training in sentence-medial positions in a pragmatically neutral context. Of course, the presence of the original pitch contour is not crucial to making a correct response with regard to the names of objects (participants make correct responses even in cases where the contours differ), nevertheless, it appears that there is a subtle sensitivity to the pitch accents presented during the training phase, despite the fact that a) these contours play no pragmatic role, b) they are all presented in the same syntactic and rhythmical context, c) German is neither a tone language nor a pitch accent language, i.e. pitch is not a distinctive feature and d) pitch accents were not employed contrastively in the training phase. Nevertheless, the subjects in our experiment attended to pitch. Hence, participants were unable to ignore a difference in the stimuli that should not be relevant at the lexical level.[2] While there is evidence that speakers attend to features that are distinctive lexical features in their language even though they are explicitly instructed to ignore them [21], the finding that they are influenced by features that are not lexically distinctive in their language appears difficult to explain. The traditional autosegmental-metrical perspective on pitch accents in German assumes that the lexical entry for a word and the pitch accent realised with it are clearly separated. Usage-based models of speech perception assume that the categorisation of incoming stimuli involves similarity comparisons to extant exemplars stored in memory in rich detail. Furthermore, such models manipulate the degree to which particular features affect categorisation by enabling different dimensions of the perceptual space to be attended to, to varying extents, via the use of attention weights [5, 22], providing a possible explanation for the results. In this experiment the exemplars in memory are representations of the audio-visual episodes encountered in training. All object names presented in the testing phase have been heard in training. Therefore, when a participant is presented with an auditory stimulus during the test phase, this stimulus matches, at least at the nominal level (and potentially at the tonal level, depending on the condition) the nominal (and potentially tonal) portion of an audio-visual representation in memory. This will lead to the activation of the entire audio-visual episode and the visual component of that memory can be compared against the image which subsequently appears on screen. If one assumes that the nominal level receives greater attention than the tonal level – German, after all, is not a tone language – then the significant latency results can be explained as follows:

[2] One might argue that pitch is somewhat relevant because it can be used to mark lexical stress, however it is not the main correlate and not used consistently across speakers in German [20].

Reaction times in the nominal mismatch condition (top line figure 2) are significantly longer than in the nominal match condition (bottom line) which can be understood in terms of greater distance in perceptual space between the image called to mind by the auditory stimulus and the image displayed on the screen. Within the nominal mismatch condition, reaction times are significantly longer if there is a tonal match than if there is a tonal mismatch. That is, a complete mismatch yields faster rejection times than a nominal mismatch but tonal match. One possible explanation is that in the case of the complete mismatch both acoustic contributors to the decision to reject (the nominal and tonal levels), are in harmony, facilitating faster rejection. Whereas in the nominal mismatch but tonal match condition, there is a conflict between the two levels which contributes to a delayed rejection decision, resulting in a longer reaction time.

It remains unclear why the delay within the nominal match condition (bottom line figure 2) is not significant. Nevertheless, the significant results within the nominal mismatch condition indicate some interaction between pitch accents and words which a strong auto-segmental perspective would not anticipate.

Concerning syllable frequency, the effect observed (faster reaction times for frequent syllables) is in line with previous findings: e.g. [15] examined the production of a stimulus associated with a position on the screen. Stimuli with a frequent first syllable were produced faster than infrequent ones. Our results extend previous findings by pairing a visual stimulus (the object) with an auditory one.

5. Conclusion

The results of this experiment offer tentative evidence of a greater cohesion between words and their tonal realisation than might traditionally have been envisaged for German. Future work will further examine the nature of this cohesion. For instance, a probabilistic assignment of pitch accents to the nonsense words in the training phase – rather than a binary one – would yield more realistic learning conditions. Furthermore, a delay between training and testing could offer insights into temporal aspects of the interaction. Finally, comparison to a similar experiment employing synthetic stimuli will also be investigated.

6. Acknowledgements

Supported by the German Research Foundation (SFB-732)

7. References

[1] J. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. dissertation, MIT, 1980.

[2] M. Liberman, "The Intonation System of English," Ph.D. dissertation, MIT, 1975.

[3] M. Jilka and B. Möbius, "The influence of vowel quality features on peak alignment," in *Proceedings of Interspeech 2007 (Antwerpen)*, 2007, pp. 2621–2624.

[4] S. D. Goldinger, "Words and voices—perception and production in an episodic lexicon," in *Talker variability in speech processing*, K. Johnson and J. W. Mullennix, Eds. San Diego: Academic Press, 1997, pp. 33–66.

[5] K. Johnson, "Speech perception without speaker normalization: An exemplar model," in *Talker Variability in Speech Processing*, K. Johnson and J. W. Mullennix, Eds. San Diego: Academic Press, 1997, pp. 145–165.

[6] J. Pierrehumbert, "Exemplar dynamics: Word frequency, lenition and contrast," in *Frequency and the Emergence of Linguistic Structure*, J. Bybee and P. Hopper, Eds. Amsterdam: Benjamins, 2001, pp. 137–157.

[7] M. Walsh, B. Möbius, T. Wade, and H. Schütze, "Multilevel exemplar theory," *Cognitive Science*, vol. 34, pp. 537–582, 2010.

[8] S. Calhoun and A. Schweitzer, "Can intonation contours be lexicalised? Implications for discourse meanings," in *Prosody and Meaning (Trends in Linguistics)*, G. Elordieta Alcibar and P. Prieto, Eds. Mouton DeGruyter, 2012.

[9] K. Schweitzer, M. Walsh, S. Calhoun, and H. Schütze, "Prosodic variability in lexical sequences: Intonation entrenches too," in *ICPhS Proc.*, Hong Kong, 2011, pp. 1778–1781.

[10] B. Braun, A. Dainora, and M. Ernestus, "An unfamiliar intonation contour slows down online speech comprehension," *Language and Cognitive Processes*, vol. 26, no. 3, pp. 350–375, 2011.

[11] B. Braun and E. K. Johnson, "Question or tone 2? How language experience and linguistic function guide pitch processing," *Journal of Phonetics*, pp. 585–594, 2011.

[12] J. M. Brenier, A. Nenkova, A. Kothari, L. Whitton, D. Beaver, and D. Jurafsky, "The (non)utility of linguistic features for predicting prominence in spontaneous speech," in *in IEEE/ACL 2006 Workshop on Spoken Language Technology*, 2006, pp. 54–57.

[13] A. Nenkova, J. Brenier, A. Kothari, S. Calhoun, L. Whitton, D. Beaver, and D. Jurafsky, "To memorize or to predict: Prominence labeling in conversational speech," in *Proceedings of NAACL-HLT*, 2007, pp. 9–16.

[14] I. Gauthier and M. Tarr, "Becoming a "greeble" expert: exploring mechanisms for face recognition." *Vision Research*, vol. 37, pp. 1673–1682, 1997.

[15] J. Cholin, W. J. M. Levelt, and N. O. Schiller, "Effects of syllable frequency in speech production," *Cognition*, vol. 99, no. 2, pp. 205–235, March 2006.

[16] K. Müller, "Probabilistic Syllable Modeling Using Unsupervised and Supervised Learning Methods," PhD thesis, University of Stuttgart, IMS, Stuttgart, 2002.

[17] J. Mayer, "Transcribing German intonation – the Stuttgart system," Universität Stuttgart, Tech. Rep., 1995.

[18] M. Tucker, "Slide Generator: A DirectX based Experiment Generator for Psychology (Version 2007.3.3)," 2007. [Online]. Available: www.psy.plymouth.ac.uk/research/mtucker/SlideGenerator.htm

[19] H. Baayen, "languageR: Data sets and functions with "Analyzing Linguistic Data: a practical introduction to statistics"," 2007, r package version 1.0.

[20] M. Jessen, K. Marasek, K. Schneider, and K. Claßen, "Acoustic correlates of word stress and the tense/lax opposition in the vowel system of german," in *Proceedings of ICPhS*, Stockholm, Sweden, 1995, pp. 428–431.

[21] E. Dupoux, C. Pallier, N. Sebastian, and J. Mehler, "A destressing deafness in french?" *Journal of Memory and Language*, vol. 36, no. 3, pp. 406 – 421, 1997.

[22] R. M. Nosofsky, "Attention, similarity, and the identification-categorization relationship," *Journal of Experimental Psychology: General*, vol. 115, no. 1, pp. 39–57, 1986.