# Pitch Accents and Information Status in a German Radio News Corpus

*Katrin Schweitzer, Arndt Riester, Michael Walsh, Grzegorz Dogil*

Institute for Natural Language Processing
University of Stuttgart

`<firstname>.<surname>@ims.uni-stuttgart.de`

## Abstract

This paper presents a corpus analysis of prosodic realisations of information status categories in terms of pitch accent types. The annotations base on a recent annotation scheme for information status [1] that is based on semantic criteria applied to written text. For each information status category, typical pitch accent realisations are identified. Moreover, the relevance of the strict semantic information status labelling scheme on the prosodic realisation is examined. It can be shown that the semantic criteria are reflected in prosody, i.e. the prosodic findings corroborate the theoretical assumptions made in the framework.

**Index Terms**: information status, information structure, givenness, discourse analysis, prosody, pitch accent types

## 1. Introduction

Intonation plays a significant role in marking information status. Even in German, with its relatively free word order, where other means of structuring the discourse (e.g. topicalisation) are very common, intonation significantly influences the interpretation of the discourse with respect to information status.

The aim of this study is, firstly, to detect tonal correlates of information status categories in a German radio news corpus, i.e. natural speech as opposed to artificial lab-speech is examined. The characteristic under investigation is pitch accent type, that is the phonological category of a tonal event evoking prominence of a syllable/word within a sentence.

Secondly, the study aims to corroborate a recent annotation scheme for information status [1], that follows strict semantic criteria and assumes categories that have not been used in previous information status annotation schemes. The goal of the phonological analysis is to determine whether these categories are reflected in pitch accent choice within spoken language.

The structure of the paper is as follows: section 2 gives an overview on information status and previous work on the correlation between pitch accents and information status, section 3 describes the database and the annotation of information status. The results are presented in section 4 and interpreted in section 5. Section 6 summarises the results and presents ideas for future work.

## 2. Background

Information status describes expressions in a discourse in terms of whether they already occurred or whether they are new to the discourse. There is a solid body of evidence that intonation is a main marker for information status.

### 2.1. Information Status

The concept of information status [2, 3] involves classifying noun phrases in texts according to various ways of their being given or new. It replaces and specifies more clearly the often vaguely used term "givenness". The process of labelling a corpus for information status can be seen as a means to facilitate discourse analysis. Different classification systems have been proposed in the literature; see [4] for a comparison of several information status labelling schemes.

### 2.2. Pitch Accents and Information Status

It is commonly assumed that pitch accents are a main correlate of information status [5]. Generally, accenting is said to signal *novelty* while deaccenting signals *given* information (e.g. [6]); yet there is counter evidence: various studies note accented information being *given* [7, 8, 9] and it has also been pointed out out that *new* information can be deaccented [10].

As for the question of which pitch accent type (in terms of ToBI categories) is typically assigned in the case of different degrees of givenness, H∗ (an accent with a high target in the accented syllable) is found to be the standard *novelty* accent for English [11] and German [12]. *Given* information, on the other hand, if accented at all, carries the L∗-accent (low target in the accented syllable) in English [11]. In German, deaccentuation is stated to be the most preferred realisation for *given* tokens in experimental studies [12]. However, [12] points out that H+L∗ (a high target followed by a low target in the accented syllable) is also found as a marker for the information status *given* in a corpus study.

Depending on the definition of *accessibility* (cf. 3.2), different preferred realisations are found. In addition to H+L∗, deaccenting was also judged acceptable in a perception experiment [12]. Analogous results are reported for a corpus analysis [12] where H+L∗ and deaccenting were the the most frequent realisations of accessibility, although L∗ has also been frequently chosen.

The study presented here analyses connections between information status categories as defined in a new taxonomy [1] and pitch accent types in terms of GToBI(S) categories. Special attention is paid to the special features of the new taxonomy and to the question whether they are reflected in prosody, or in pitch accent choice, more precisely.

## 3. Methodology

For this study, a German radio news corpus was analysed (cf. section 3.1). The textual transcription of the corpus was annotated with respect to information status, independently of any prosodic information (section 3.2). The noun constituents that were annotated with information status were examined for syllables that are likely to be the exponent of information status (cf. section 3.3).

6 – 10 September, Brighton UK

## 3.1. Data

The experiment was carried out on data from the IMS Radio News Corpus [13]. This corpus is automatically segmented and manually labelled according to GToBI(S) [14] and consists of approximately 1hour of speech. It contains data from three speakers, two female and one male.

The orthographic transcription of the speech signals was manually labelled with respect to information status. As the annotation scheme is based solely on the written text (cf. section 3.2) which might be underspecified in terms of information status category, it allows for multiple information status labels on the same noun constituent.

To draw clear-cut conclusions on pitch accent preference of the different information status categories, a one-to-one relationship between information status category and tonal properties is required. Therefore, only a subset of the corpus was used: it includes only syllables to which just one information status category has been applied.

## 3.2. Annotation of Information Status

The annotation of information status is based on a semantically motivated taxonomy [1]. The main theoretic assumption is that information status categories (for definites) should group expressions according to the contextual resources in which their antecedents are found (e.g. "next Monday" is resolved in the situative, "Angela Merkel" in the encyclopaedic context).[1] Based on this, several information status categories are assumed. Those that are relevant for the corpus used in this study are described as follows:

**d-given-*:** The group of *d-given* categories is defined on the basis of coreference. Expressions that refer to entities that are given in the **d**iscourse context fall under that criterion. The scheme offers a more fine grained distinction between different types of d-givenness: *d-given-repeated* for exact repetitions, *d-given-epithet* for expressions that use new lexical material to refer to an entity that has already been mentioned in the discourse, *d-given-pronoun* for pronominals and *d-given-short* for short forms of expressions that have already been mentioned. Different classes of *d-givenness* can be found in the following example:

(1)  [Angela Merkel]$_{\text{ACC-GEN}}$ (*first mention*) . . .
[Angela Merkel]$_{\text{D-GIV-REPEATED}}$ (*second mention*) . . .
[Merkel]$_{\text{D-GIV-SHORT}}$ · · · [she]$_{\text{D-GIV-PRONOUN}}$ · · ·
[the Hamburg-born politician]$_{\text{D-GIV-EPITHET}}$

**accessible-general:** Expressions that are not present in the previous discourse but refer to entities known to the intended recipient and can thus be resolved in the encyclopaedic (also: knowledge) context (cf. example (1), first expression).

**situative:** Expressions referring to antecedents in the situative context (i.e. the speaker environment). Typically these are the discourse participants, demonstratives referring to objects in the speaker situation or deictic expressions, as below:

(2)  this setback needs to be counteracted [now]$_{\text{SITUATIVE}}$

**accessible-description:** Expressions that are not resolvable in any of the contexts (they have not been mentioned, nor does the audience know or perceive them) and hence, have to be accommodated, as in example (3). This is an issue that has not

been addressed in earlier annotation schemes, in which these expressions would have most likely been subsumed under the concept of *general accessibility*.

(3)  [the leadership crisis lasting for months among the Hamburg Social Democrats]$_{\text{ACC-DESC}}$

**new:** Indefinite expressions that are unrelated to context as in the following example:

(4)  the peace talks have been continued in spite of [a deep crisis]$_{\text{NEW}}$

The main differences to other information status taxonomies are the newly introduced category *accessible-description*, the definition of givenness based on coreference (in contrast to literal previous mention, as it is defined e.g. in [12], following [15] and [16]), the fine-grained distinction into several different classes of givenness and a relatively tight definition of novelty which, unlike other taxonomies, does not include definites.

The annotations were conducted by two trained independent annotators and subsequently compared by a third person who – in cases of disagreement – took the ultimate decision.

## 3.3. Extracting the Relevant Syllables

The annotation scheme attaches information status labels to full noun phrases rather than to the nouns themselves. These constituents can be very long and can therefore not only include several potentially accentable syllables, but also several pitch accents. It is thus not trivial to decide, which syllable(s) of a noun constituent should be included in the qualitative analysis.

Including all potentially accentable syllables (those that carry word stress) results in an over-proportion of non-accented syllables and does not reflect reality. Only including nuclear accents, on the other hand, leads to an analysis that has no means to measure deaccentuation. In the present study, a middle course was chosen: the analysis includes only those syllables that carry word stress (and are thus potentially pitch-accentable) and that are part of the last word in the annotated noun constituent. Thereby the word stressed syllable of the constituent's head is defined as the relevant syllable in most of the cases. These restrictions apply to 608 syllables in the corpus.[2]

# 4. Results

The following section describes the methodology of statistical testing (section 4.1) and reports the results for the different information status categories (section 4.2). Special attention is paid to the categories that are new in the taxonomy of [1] and have not been analysed before: firstly, the category *accessible-description* in comparison to the category *accessible-general* under which the respective expressions would have been subsumed in other taxonomies and secondly the categories for *discourse givenness* in general, as well as the fine-grained distinction into four different types.

## 4.1. $\chi^2$-Test for Independence and Sparse Data Problems

To test for potential dependency of information status category and pitch accent type, $\chi^2$-tests for independence of the two vari-

---

[1]This stands in contrast to several other annotation schemes following [15] in which the categories are mainly grouped according to cognitive activation.

[2]N.B. In those cases where the syllables are accented (382), nuclear as well as non-nuclear accents are analysed. Including both types yields an adequate number of syllables, but does not take the effects of prosodic context into account (cf. section 6). However, nuclear accents occur more than twice as often (273) as non-nuclear ones (109).

ables were applied. Table 1 shows the contingency table reflecting the relationship between information status categories and pitch accent types. The $\chi^2$-test resulted in a significant p-value of $p \ll 0.00001$ indicating a dependency between the two factors. However, it has to be noted that the $\chi^2$ approximation might be incorrect because of the lack of data in several combinations of information status and pitch accent type.

|          | H* | HH*L | !H*L | H*L | L*HL | L*H | L*!H | NONE |
|----------|-----|------|------|-----|------|-----|------|------|
| acc-descr | 3  | 1    | 4    | 10  | 1    | 13  | 1    | 47   |
| gen.-acc  | 1  | 2    | 2    | 18  | 2    | 52  | 2    | 34   |
| situative | 4  | 0    | 2    | 10  | 0    | 43  | 4    | 33   |
| d-giv-pro | 2  | 0    | 0    | 0   | 0    | 10  | 0    | 29   |
| d-giv-rep | 0  | 0    | 0    | 2   | 0    | 8   | 0    | 7    |
| d-giv-short | 1 | 0   | 2    | 1   | 1    | 9   | 1    | 6    |
| d-giv-epi. | 3 | 1    | 1    | 14  | 1    | 39  | 2    | 32   |
| new       | 4  | 1    | 12   | 40  | 5    | 37  | 10   | 38   |

Table 1: Frequency of pitch accent types accross information status categories

Therefore, a more coarse-grained testing procedure was applied in which pitch accent types were subsumed under three different categories: *falling* for the falling accents H*L, HH*L, H*, and L*HL, *rising* for the rising accent L*H and its downstepped version L*!H, and *NONE* for unaccented syllables. With these three categories all cells of the contingency table are filled and in only two cases the frequency is less than 5 tokens (see table 2). The $\chi^2$-test also results in a significant p-value of $p \ll 0.00001$.

|          | falling | rising | NONE |
|----------|---------|--------|------|
| acc-descr. | 19    | 14     | 47   |
| gen.-acc  | 25      | 54     | 34   |
| situative | 16      | 47     | 33   |
| d-giv-pro | 2       | 10     | 29   |
| d-giv-rep | 2       | 8      | 7    |
| d-giv-short | 5     | 10     | 6    |
| d-giv-epi. | 20     | 41     | 32   |
| new       | 62      | 47     | 38   |

Table 2: Frequency of coarse-grained pitch accent categories accross information status categories

In the following, all tests that are reported as significant were applied to both the fine-grained pitch accent distinction as well as the coarse-grained distinction that avoids sparse data problems. In all cases, significance holds across both tests.

### 4.2. Pitch Accents and Information Status in the IMS Radio News Corpus

Table 3 displays the relative frequency in percentage of each pitch accent type within an information status category.

It clearly displays the difference between the two information status categories *accessible-description* and *accessible-general*. While the former is mainly realised without an accent (58.75%), the latter shows a clear preference for the rising accents (L*H and L*!H), which add up to 47.79% of all cases.

The difference in pitch accent choice between these two information status categories is statistically significant ($p < 0.005$).

*Situatives* are mainly realised with a rising accent (L*H and L*!H add up to 48.96%). The second most preferred realisation is deaccentuation which occurs in 34.38% of all cases.

Among the subcategories of *d-givenness*, pitch accent type is significantly dependent on information status ($p < 0.05$). A closer look gives insight into the nature of the differences: while *d-given-pronoun*, *d-given-epithet* and *d-given-repeated* do not differ significantly in pitch accent choice, the category for pronouns, *d-given-pronoun*, differs significantly from the others ($p < 0.005$). This is clearly due to the fact that pronouns are deaccented in the vast majority of the cases (70,73%) while the other categories of *d-givenness* prefer rising accents: the relative frequency ranges between 44.05% for *d-given-epithet* and 47.62% for *d-given-short*. Deaccentuation is also very common for all three categories (28.57% for *d-given-short*, 34.41% for *d-given-epithet* and 41.18% for *d.given.repetition*). However, it should be noted that for the categories *d-given-short* and *d-given-repeated* the number of tokens is relatively small (21 and 17 tokens, respectively).

In contexts where the label *new* had been applied, no clear preference for a pitch accent type can be found, H*L, L*H and deaccentuation are almost evenly distributed, though falling accents in general (occurring in 42.18% of the cases) are preferred over rising accents (31.97%) or deaccentuation (25.85%).

|          |      | falling |      |      |      | rising |      |      |
|----------|------|---------|------|------|------|--------|------|------|
|          | H*   | HH*L    | !H*L | H*L  | L*HL | L*H    | L*!H | NONE |
| acc-descr. | 3.75 | 1.25   | 5.00 | 12.50 | 1.25 | 16.25 | 1.25 | **58.75** |
| gen.-acc  | 0.88 | 1.77   | 1.77 | 15.93 | 1.77 | **46.02** | **1.77** | 30.09 |
| situative | 4.17 | 0.00   | 2.08 | 10.42 | 0.00 | **44.79** | **4.17** | 34.38 |
| d-giv-pro | 4.88 | 0.00   | 0.00 | 0.00 | 0.00 | 24.39 | 0.00 | **70.73** |
| d-giv-rep | 0.00 | 0.00   | 0.00 | 11.76 | 0.00 | **47.06** | 0.00 | 41.18 |
| d-giv-short | 4.76 | 0.00 | 9.52 | 4.76 | 4.76 | **42.86** | **4.76** | 28.57 |
| d-giv-epi. | 3.23 | 1.08   | 1.08 | 15.05 | 1.08 | **41.94** | **2.15** | 34.41 |
| new       | **2.72** | **0.68** | **8.16** | **27.21** | **3.40** | 25.17 | 6.80 | 25.85 |

Table 3: Relative frequency of pitch accent types within each information status category. Preferred pitch accent types are bold-faced for each information status category.

## 5. Interpretation

In the current study the category *accessible-description* has been found to significantly differ from the category *accessible-general* under which the respective entities would have been subsumed in former information status labelling schemes. While *accessible-general* items are predominantly realised with rising accents, *accessible-description* items normally do not receive a pitch accent. This represents evidence that the semantically motivated category is significant for the prosodic realisation of information status. Together, the results for accessibility are in keeping with other findings for German [12], though a comparison is not trivial as the definition for accessibility as well as the intonation labelling systems differ. But in both studies either a preference for an accent with a low target tone on the accented syllable (L*H here, H+L* in [12]) or deaccentuation is found as a marker for accessibility.

*Situatives* were found to prefer rising accents, as *accessible-general* entities do. That is, in the analysed data, the difference

between situative and encyclopaedic context is not reflected in prosody. [12] also reports H+L∗ and deaccentuation for these expressions, which is consistent with the findings reported here (accents with a low target as the most preferred realisation pattern, deaccentuation ranking second).

The fine-grained distinction into four different types of *discourse-givenness* is only reflected in prosody for the category for pronouns, *d-given-pronoun*, that is predominantly not accented – given the general prominence of pronouns and their shortness this seems intuitive. Generally it can be said that, in the corpus, the realisation of *discourse givenness* is similarly to the realisation of *accessible-general* and *situative* items with rising accents being the most preferred pitch accent types followed by deaccentuation. Thus, the semantic difference between co-reference (in the case of *d-givenness*, e.g. "the Hamburg born politician") and resolution in the encyclopaedic or situative context (in the case of *situative* e.g. "next Monday" and *accessible-general* e.g. "Angela Merkel") is not reflected in prosody – at least not in terms of pitch accent categories. In comparison to the findings of [12] it is noteworthy that *discourse-givenness* in general and givenness as 'literal previous mention' (as it is defined in [12]) are realised similarly, though the two definitions cover different expressions. This highlights the importance of both aspects for a global definition of givenness.

*Novelty* in the current study, is not clearly marked by one particular pitch accent though falling accents are preferred slightly over rising accents and deaccentuation. This signals the special status of indefinites. Moreover it leads to the assumption that for *new* expressions other factors (i.e. position in the phrase) have a significant influence on the choice of pitch accent type. The over-proportion of falling accents (i.e. accents with a high target in the accented syllable) is consistent with [12] who reports H∗ to be the typical indicator for novelty, but the results will have to be examined more closely in terms of the different ToBI-labelling systems for German as well as with respect to different definitions of novelty: entities which are classified as *accessible-general* in this study, would have also been categorised as *new* according to [12].

## 6. Conclusion and Outlook

In the current study, typical pitch accent types (including deaccentuation) for information status categories as defined in [1] were identified. While some categories (*accessible-description* and *d-given-pronoun*) show a clear preference for one particular realisation, others additionally show a tendency for a second-best candidate (*accessible-general* and *situative*, as well as the other types of *d-givenness*). Interestingly, all these categories are predominantly realised with a rising accent but also show a tendency for deaccentuation. The category *new* is the only category for which the different accents types are almost evenly distributed across the analysed expressions, which might hint at a special status of indefinites.

The results highlight the importance of semantic criteria for an all-encompassing definition of information status. The special category for anaphora that have to be accommodated in the context, *accessible-description*, is realised significantly differently from the category *accessible-general*, that would subsume the respective entities in previous systems. The subdivision into different types of *discourse-givenness* is significant for pitch accent choice in the case of givenness realised as a pronoun, where deaccentuation is the preferred realisation. The other types of givenness behave similarly, reflecting the fact that they are all

based on co-reference. The fact that *discourse-givenness* behaves similar to givenness in terms of 'literal previous mention' [12] shows that there are semantic as well as textual criteria that define the same concept.

The intricate relationship between pitch accents and information status has to some extent been further elucidated in this study. Further opportunities for investigations remain: for example, in the study nuclear and non-nuclear accents have been analysed together for the sake of a greater amount of data. In ongoing research, the prosodic context of a syllable is examined more closely, e. g. nuclear vs. non-nuclear contexts are under investigation.

## 7. References

[1] A. Riester, "The components of focus and their use in annotating information structure," Ph.D. dissertation, University of Stuttgart, 2008.

[2] E. F. Prince, "Toward a Taxonomy of Given-New Information," in *Radical Pragmatics*, P. Cole, Ed. New York: Academic Press, 1981, pp. 233–255.

[3] ——, "The ZPG Letter: Subjects, Definiteness and Information Status," in *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, W. C. Mann and S. A. Thompson, Eds. Amsterdam: Benjamins, 1992, pp. 295–325.

[4] A. Riester, "A Semantic Explication of Information Status and the Underspecification of the Recipients' Knowledge," in *Proceedings of Sinn und Bedeutung 12*, A. Grønn, Ed., Oslo, 2008.

[5] M. A. K. Halliday, *Intonation and Grammar in British English*. The Hague: Mouton, 1967.

[6] G. Brown, "Prosodic structure and the given/new distinction," in *Prosody: Models and Measurements*, A. Cutler and D. R. Ladd, Eds. New York: Springer, 1983, pp. 67–77.

[7] S. G. Nooteboom and J. G. Kruyt, "Accents, focus distribution, and the perceived distribution of given and new information: an experiment," *Journal of the Acoustical Society of America*, vol. 82, pp. 1512–1524, 1987.

[8] G. Yule, "Intonation and Givenness in Spoken Discourse," *Studies in Language*, pp. 271–286, 1980.

[9] E. G. Bard and M. P. Aylett, "The dissociation of deaccenting, givenness, and syntactic role in spontaneous speech," in *Proceedings of the 14th International Congress of Phonetic Sciences (San Francisco)*, vol. 3, 1999, pp. 1753–1756.

[10] J. Terken and J. Hirschberg, "Deaccentuation of words representing 'given' information: effects of persistence of grammatical function and surface position," *Language and Speech*, vol. 37, pp. 125–145, 1994.

[11] J. Pierrehumbert and J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in Communication*, P. R. Cohen, J. Morgan, and M. E. Pollack, Eds. Cambridge, MA: MIT Press, 1990, pp. 271–311.

[12] S. Baumann, *The Intonation of Givenness – Evidence from German.*, ser. Linguistische Arbeiten. Tübingen: Niemeyer, 2006, vol. 508, Ph.D. thesis, Saarland University.

[13] S. Rapp, "Automatisierte Erstellung von Korpora für die Prosodieforschung," Ph.D. dissertation, IMS, Universität Stuttgart, 1998, aIMS 4 (1).

[14] J. Mayer, "Transcribing German Intonation – The Stuttgart System," Universität Stuttgart, Tech. Rep., 1995, http://www.ims.uni-stuttgart.de/phonetik/joerg/labman/STGTsystem.html.

[15] W. L. Chafe, *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University Of Chicago Press, 1994.

[16] K. Lambrecht, *Information Structure and Sentence Form*. Cambridge: Cambridge University Press, 1994.