INFORMATION STRUCTURE IN SPEECH SYNTHESIS: EARLY FOCUS AND POST-FOCAL GIVENNESS

Frank Kügler, Bernadett Smolibocki, Manfred Stede and Sebastian Varges
Dept. of Linguistics/EB Cognitive Science & SFB 632 "Information structure"
University of Potsdam
kuegler@uni-potsdam.de

Abstract: Even though speech synthesis nowadays is of acceptable quality for many purposes, straightforward text-to-speech (TTS) systems do not produce optimal results in cases where contextual and other pragmatic factors play an important role for prosodic realization. For instance, in systems giving product comparisons and recommendations, an appropriate intonation is required to signal contrasting entities; and in longer discourse, given and new entities need to be distinguished prosodically. In our project, such notions of information structure (IS) are used to extend an existing text generator for product comparison/recommendation with a speech synthesis component (MARY TTS). In this paper, we concentrate on one particular IS phenomenon: *post-focal givenness*. The purpose of the paper is twofold: First, we explain the architecture of our system and the IS extensions we made MARY TTS (MARY+IS); second, we show that an appropriate prosodic marking of post-focal givenness indeed leads to increased hearer acceptability ratings.

1. Introduction

Over the past years, domain- and task-independent text-to-speech synthesis (TTS) has reached an impressive quality, which is sufficient for a range of practical applications. In some speech production scenarios, however, pragmatic factors play an important role: The goals of the speaker can have consequences for appropriate prosodic realization, or the linguistic context may impose preferences on various prosodic aspects, or both. In our work, we specifically deal with aspects of information structure (IS), where the preceding context of an utterance has ramifications for the "information packaging" [5] of the current utterance. In Section 2, we briefly explain what particular notions of IS we use.

Our application scenario is product comparison and recommendation: In response to a user's query about mobile phones (the current domain of choice) that suit his or her needs well, the system peruses a large database of devices to select a few that fit the user's description. Next, it generates text that compares phones to one another. Intuitively, it is clear that this is a setting where IS is indeed of relevance: The same objects are being talked about multiple times, their identical or similar attributes are compared and possibly contrasted. In Section 3, we describe the scenario for the text generator and its handling of the IS features. In particular, for the sake of illustration, we will discuss the production of a single sentence (*Das iPhone funktioniert im Netz von T-Mobile.* 'The iPhone functions in the net of T-mobile.'), which will later on be discussed in the report of our experiments.

Our main technical goal on the speech synthesis side is to produce extensions to the MARY synthesizer [25] in order to enable it to systematically respond to IS-related markup of the input text. In effect, this is a move from "plain" TTS to annotated-text-to-speech, which may be seen as a step toward the idea of concept-to-speech synthesis (CTS). We have defined a set of IS-tags that are to be used for this markup, and we started to implement the necessary extensions to MARY TTS, i.e. MARY+IS. For the purposes of this paper, we focus on enhancing the system to deal with post-focal givenness. In Section 4, we report on experiments that confirm the hypothesis that listeners indeed prefer the "IS aware" output to the standard output of MARY TTS. Finally, Section 5 draws conclusions.

2. Information structure

It is a truism that the appropriate information packaging of an utterance depends on features of the preceding discourse. However, linguistic approaches differ widely on the issue of how to spell out the various dimensions of information structure (IS), and how to relate them to specific features of linguistic utterances. For our work, we follow the proposal of [14], who broadly distinguishes between the dimensions *given/new, topic/comment,* and *focus/background*, and offers several more fine-grained distinctions for these categories. For the linguistic realization of the dimensions, languages differ according to their inventory of syntactic and prosodic means; for our purposes here, we concentrate on prosodic features as they are used in intonation languages such as German [7,9,10,13,17].

This paper focuses on the information status of *post-focal given* constituents. Givenness, which for German was studied extensively by [3], is commonly defined in terms of a referent having been mentioned in the previous discourse, or in [14]'s slightly different description, "the denotation of an expression is present in the immediate common ground content." Several authors have proposed subclassifications of *degrees* of givenness. We follow [11] in differentiating between referents that are *active* (mentioned in the last or in the current sentence), *inactive* (mentioned before the last sentence), or *accessible* (not mentioned before but prominent due to a relation with a mentioned referent, or due to world knowledge). Turning to prosodic realization, it has been argued that for given constituents in post-focal position, a universal means for signaling their IS status is to not use accents [6:313]. In order to operationalize this insight for MARY TTS, we will in Section 4 suggest a variant of this strategy by means of deaccentuation.

Another IS dimension that is relevant for our present goals is *contrast*. One pragmatic use of contrast is the correction of a previously mentioned piece of information. We view this *corrective* focus as a subtype of contrastive focus [11,23]. Relevant phonetic cues of corrective focus concern the horizontal and vertical alignment of a pitch peak. For example, corrective focus is realized by an enhanced pitch register [9,16]. Consider example (1) below, where in the answer *iPhone* corrects the item *Nokia Lumia* of the question.

(1) a) **Context question**
Funktioniert das Nokia Lumia im Netz von T-Mobile?
function the Nokia Lumia in.the net of T-mobile

b) **Answer**
Das iPhone funktioniert im Netz von T-Mobile.
the iPhone function in.the net of T-mobile

3. Text and Speech Generation System

The application serving as the background to our research is the product recommendation system *Polibox* [26], which has been extended to operate in the domain of mobile phones. The user explicitly states his or her needs by providing some target product features such as price, weight, camera quality, etc. The system then suggests products, compares and possibly actively recommends them. In contrast to many competing approaches that merely generate tabular output, Polibox produces natural language texts.

The underlying architecture involves a classical language generation pipeline [22] involving content selection, document planning, sentence planning, and syntactic realization. We use OpenCCG [21], equipped with a grammar for German, as the realization engine. The speech synthesis front-end is MARY [25], which we have started to enhance with features for IS processing, and thus henceforth will call it MARY+IS. When the generator plans a text, it keeps track of the entities being talked about in its discourse memory. Moreover, the semantic input of the realizer is being enriched with IS features, following the tagset of [11]. Let us illustrate this with the example (1) above where we want the system to generate the answer (1b) with “iPhone” labeled by a “corrective-contrastive” marker. The underlying propositions selected by the system to generate (1b) are given in (2) (in slightly simplified format):

(2) Propositions: 1:model(ref1, iphone). 2:network_availability(ref1, t-mobile).

In proposition 1, ‘ref1’ refers to the referent ‘iPhone’ in the discourse memory. Proposition 2 expresses a fact about this referent. In addition, the information state of the propositions is determined:

(3) Information State: 3:is_label(1, contrast). 4:is_label(2, given-active).

An ‘is_label’ states meta-information about a proposition: proposition 1 forms a corrective contrast (to ‘Nokia Lumia’) but the network information has already been given and thus carries the label ‘given-active’. Propositional and IS-information is passed to a pattern matching process that maps it to a sentence-semantic specification in Hybrid Logic Dependency Structure format [1,15]:

```
(4) @s1(to-function^<CTYPE>dcl^<TENSE>pres^
      <Actor>(p1^iPhone^<NUM>sg^<Det>(t1^the)^ <INFOSTAT>contrast)^
      <Location>(n1^network^<NUM>sg^<Owner>(t1^t-mobile)^
                                    <INFOSTAT>given-active))
```

The HLDS specification, which combines syntactic and semantic features with IS labels, serves as input to the openCCG realizer. In the present implementation, the IS-labels are only passed through the grammatical derivation and do not constrain it. In the final step of language generation, the syntactic structure is used to determine the application of IS-labels to words (5). As (5) shows, the IS-labels are mapped to markup suitable for the prosody module of the speech synthesizer MARY+IS:

```
(5) <maryxml> <t>das</t> <t contrast="+">iPhone</t>
              <t given="+">funktioniert</t> <t>im</t> <t given="+">Netz</t>
              <t>von</t> <t given="+">T-Mobile</t> </maryxml>
```

4. Post-focal givenness in MARY – a perception study

The underlying assumption of the speech synthesis system MARY is that a default intonation pattern of German is mapped onto a string of words. Therefore, MARY uses primarily part-of-speech information in order to assign pitch accents. Information status of discourse referents is not used because a reliable determination of information structure categories is not possible for a TTS. Hence, a neutral intonation pattern is applied, which includes several prenuclear rising and nuclear falling accents. The phonological analysis of such a pattern refers to works of [8,29] and the GToBI conventions, most recently given in [12]. Our example (1) is prosodically analyzed in MARY as in (6):

```
(6)   L+H*                  L+H*        H*   L%
      Das iPhone funktioniert im Netz von T-Mobile.
```

In an earlier version of MARY an information structural component was added that refers to the information status (lexical/semantic givenness and contrast) of discourse referents [24] since given referents are prosodically marked by means of deaccentuation [see 3,6,19, 20:174ff], whereas contrastive referents are signaled by more prominence [9,16]. Therefore, rules of accent assignment were implemented stating that a given discourse referent must not receive an accent (7a), whereas a contrastive referent must receive one (7b). However, according to [24] an appropriate prosodic realization with respect to contrastiveness as well as givenness was not implemented after all.

```
(7) a)  <attributes given="+"></attributes><action accent=""> </action>
    b)  <attributes contrast="+"></attributes><action accent="tone"></action>
```

In a recent evaluation of the prosody of the MARY synthesis, [18] found that a particular prosodic implementation of a contrastive falling accent receives higher acceptance than the current default prosody synthesis. In [18] a new accent to the prosody module of MARY+IS-1 was added in order to achieve the perceptually strong impression of a falling accent in a contrastive context. This accent was specified by a list of tuples, where the first element indicated the temporal level and the second the Hertz level. The values for the individual tuples complied with the findings of the phonetic realizations of contrastive focus [9,16]. More precisely, this means that the pitch peak of the contrastive accent, which aligned at the end of the stressed syllable of the nuclear accent, was around 25 Hz higher than the pitch peak of the default H* accent. The fall began immediately after achieving the pitch peak and ends up at around 60% Hz lower on the following word or rather at the end of the sentence. In the [18] study, the German HMM-voice *dfki-pavoque-neutral* was used. The required *contrast*-attribute of the constituent, which is considered for the following pitch accent assignments, was provided by the speech generation system (see section 3).

In [18] the position of the contrast was varied from ‘late’, via ‘medial’ to ‘early’. For late and medial occurrence of contrast higher acceptance rates were found than for early occurring contrast. The reason for this asymmetry was presumably that only one prosodic parameter was adjusted according to the information structure, i.e. the phonetic shape of the falling accent as an expression for contrastiveness. At the same time, an early occurring nuclear accent requires deaccentuation of subsequent post-focal given constituents. To a certain extent this deaccentuation was achieved by applying rule (7a). Although the prosodic information for the synthesis system contained no pitch accents in the post-focal domain slight prominences were still produced. Impressionistic tests with longer sentences than the ones tested in [18] revealed even more post-focal prominences as shown by pitch movements on given elements in Figure 1.

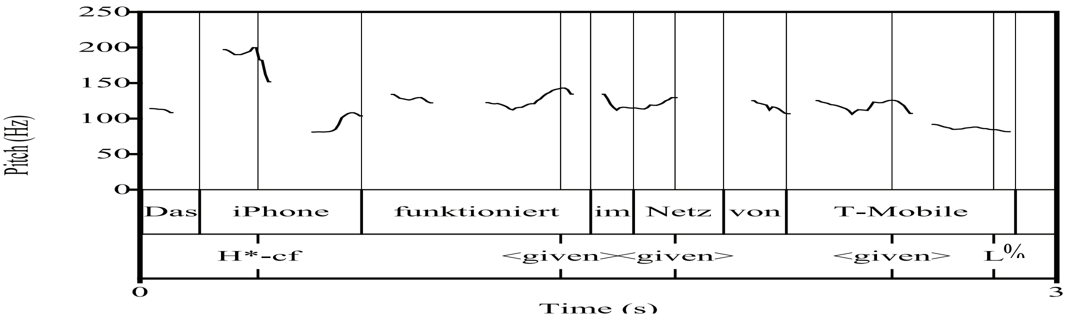

Figure 1 - F0 contour of the MARY+IS-1 synthesis of the sentence “Das iPhone funktioniert im Netz von T-Mobile” produced in a subject contrast context. The H*-cf label represents the contrast accent. The post-focal domain ranges from the beginning of the verb to the end of the sentence. The movements of high and low tones on the post-focal arguments are perceived as prominences.

The aim of the present perception study, therefore, is to extend the MARY+IS-1 version including a further reduction of post-focal prominence. In particular, MARY+IS-2 includes post-focal pitch manipulation which relies on both accent manipulation and pitch register manipulation. The hypothesis is that stimuli created with MARY+IS-2 yield significantly more congruent ratings in perception since both the contrastive accent shape of MARY+IS-1 and the deaccentuation of post-focal constituents were adjusted in the prosody module.

4.1 Speech material

Four short and four long sentences were used (see Appendix). The syntactic structure of the target sentences was S V O. The long sentences contained one more argument after the object. The added argument was either of the form of a genitive NP, or a prepositional phrase. Either of these additions contained a further referential expression that carried a pitch accent under neutral conditions. Hence, in case of deaccentuation any discourse referent needed to be realized without any pitch accent [6,20]. The manipulation of sentence length was thus between one and two referents. Example (1) above represents a long sentence. The target sentence was combined with appropriate context questions (see Appendix). The questions were produced by a human speaker eliciting contrast on the subject of the target sentence. In terms of information status the corrected constituent represented new information while the other constituents represented given information. Subject focus referred thus to sentence-initial focus.

4.2 Stimulus creation

For contrastive accents MARY+IS-2 relies on MARY+IS-1 [18]. To achieve post-focal deaccentuation two steps were implemented in MARY+IS-2. First, the pitch register of the post-focal domain was compressed at around 25Hz in order to receive a lower and more grounded baseline for the whole pitch contour. However, the post-focal arguments still revealed prominences (see Figure 1). To avoid these post-focal prominences, we secondly introduced a post-focal pitch accent TPF* (Tone_POST-FOCAL*) with certain phonetic characteristics as a technical means that meets the criteria of post-focal deaccentuation. This accent allows us to flatten the pitch contour of the arguments, and hence to reduce post-focal prominence. However, the different segmental makeup of the arguments in the different stimuli seemed to cause a great degree of microprosodic variation (for HMM-synthesis see [28]). Therefore, in the present study the specification of this accent varied slightly between the stimuli: for each of the eight stimuli a particular TFP configuration was used. Figure 2 represents the pitch contour of the same sentence as in Figure 1 but was produced by MARY+IS-2.

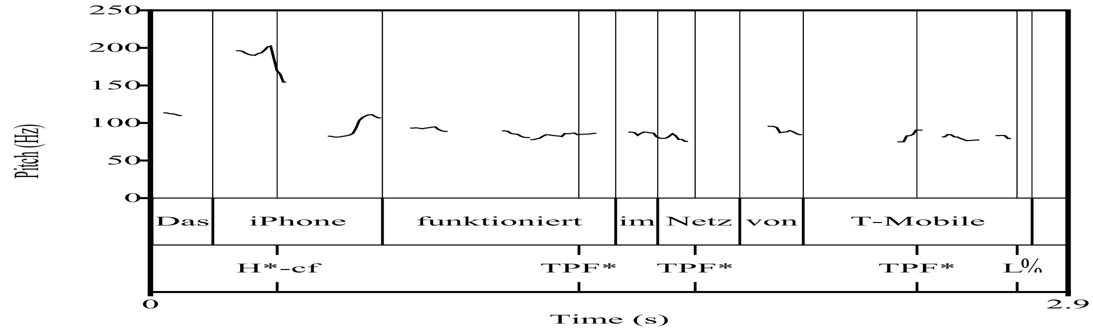

Figure 2 - F0 contour of the MARY+IS-2 synthesis of the sentence *Das iPhone funktioniert im Netz von T-Mobile.* produced in a subject contrast context. The H*-cf label represents the contrast accent, the TPF*-labels the post-focal given constituents showing deaccentuation. The post-focal domain ranges from the beginning of the verb to the end of the sentence.

4.3 Perception study

4.3.1 Listeners

21 (9 female, 12 male) native speakers of Standard German in their twenties participated in the task, all of which were graduate or undergraduate university students, and born, raised and educated in and around Berlin. None of the speakers reported any speaking or hearing deficits. All speakers were paid a small fee for participation.

4.3.2 Task

In a forced-choice semantic congruency task pairs of question and answers were presented using the Praat MFC environment [4]. Participants were placed in front of a computer screen and listened to the context question of a human voice which was followed by a MARY+IS synthesized answer. Their task was to rate the congruency of the answer to the question. Two possibilities for answering were given on the screen; (i) *match* or (ii) *no-match*. The participants had to click either on the *match*-button or on the *no-match* button after every stimulus pair. The test started after a short introduction by the experimenter and three training dialogs. The experiment lasted about 10 minutes. Overall, 32 stimuli pairs were presented (2 sentence lengths x 2 MARY+IS versions x 4 items x 2 repetitions).

4.4 Results

Figure 3 displays the congruency ratings for the previous MARY+IS-1 (leftmost two bars) and the new MARY+IS-2 (rightmost two bars). The results show that listeners rated the MARY+IS-2 version as more congruent to the context (in 74.7 %) than the previous MARY+IS-1 version (in 43.5%). Sentence length does not appear to differ in the MARY+IS-2 version. On average, the short sentences were rated 73.8 % as congruent in contrast to 75.6 % for long sentences. For the previous MARY+IS-1 version a slight difference in the congruency rating is observed. Shorter sentences were rated as congruent on average of 42.3 % in contrast to 44.6 % for longer sentences.

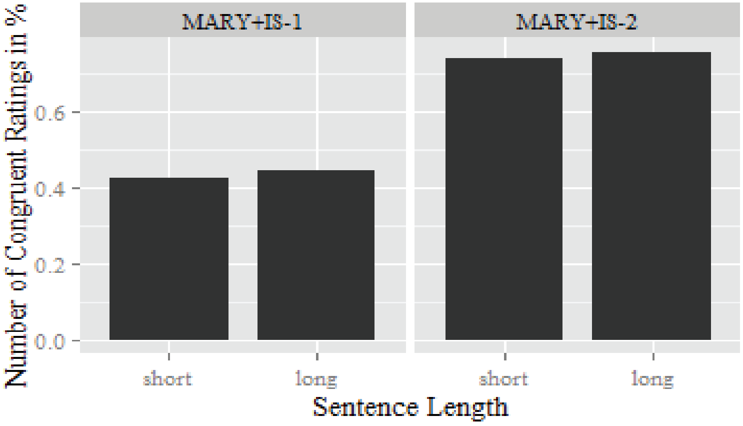

Figure 3 - Comparison of MARY+IS-1 and MARY+IS-2 in the context of initial subject contrast with post-focal de-accentuation and different sentence length, n=168 per condition

Fitting a linear mixed model [2] with both MARY+IS versions and sentence length as fixed factors and speaker and item as random factors, confirms the hypothesis that MARY+IS-2 version stimuli are rated as significantly more congruent than MARY+IS-1 version stimuli (SE 0.2656, $z = 5.784$, $p < 0.000$). Sentence length does not differ significantly (SE 0.2490, $z = 0.445$, $p = 0.657$). No significant interaction between MARY+IS version and sentence length was found (SE 0.374749, $z = -0.011$, $p = 0.991$).

4.5 Discussion

The results show that the stimuli of MARY+IS-2 were rated as more congruent than the ones of MARY+IS-1. This means that the listeners prefer the synthesis of post-focal givenness by MARY+IS-2 over the realization by MARY+IS-1. Thus, the ratings correspond to the highest one in the previous study [18]. For the MARY+IS-2 synthesis two components were considered: (i) phonetics of post-focal accents of the individual arguments and (ii) compression of the pitch register [30]. The former is contrary to the assumptions of [6,20], for example, who propose that all accents disappear on post-focal constituents. However, technically, we created a post-focal pitch accent TPF* in order to control for the deaccentuation of the arguments. With such an accent we achieve that local pitch movements of the default prosody are eliminated. Our approach yields an appropriate prosodic synthesis with respect to the information structure categories which are provided by the speech generation system.

5. Conclusion

For product comparison and recommendation systems, context dependent prosodic realizations are needed. For the synthesis of appropriate intonation contours, the system provides text that contains IS-tags as in (5) calculated from the discourse memory. These serve as input for the synthesis component that is equipped with well-defined phonetic properties of pitch accents and prosodic structure in relation to the IS-tags. Hence, the information status of discourse referents provides the basis for an IS sensitive synthesis of intonation contours. In particular, the aim of this study was to investigate the improvement of prosodic realization of post-focal givenness by MARY+IS-2. Here, post-focal deaccentuation was implemented by means of a reduced pitch register and a post-focal tone (TPF*) that controls the deaccentuation of post-focal constituents. From the perception test we can conclude that the higher congruent ratings for MARY+IS-2 mirror the importance of considering all constituents at different levels of information status [14,11]. Using a newly introduced TPF*accent, post-focal givenness of arguments after an early focus was realized as deaccentuation. The next step is to control for microprosodic influences of post-focal constituents to establish a reliable phonetic set of TFP* properties. This will serve the ultimate goal of the project to integrate IS into speech synthesis in a systematic way.

Acknowledgements

This paper is part of the project T2 of the SFB 632 on Information Structure at Potsdam University, financed by the DFG. We are grateful to Rike Schlüter for assistance.

References

[1] Baldridge, J. and G.-J. M. Kruijff: Coupling CCG and Hybrid Logic Dependency Semantics. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 2002, pp. 319-326.

[2] Bates, D. and D. Sarkar: lme4: Linear mixed-effects models using s4 classes: R package version 0.9975-11, 2007 [Computer software].

[3] Baumann, S.: The Intonation of Givenness - Evidence from German. Linguistische Arbeiten 508, Niemeyer: Tübingen, 2006.

[4] Boersma, P. and D. Weenink: Praat-doing phonetics by computer. http://www.praat.org, January 31, 2013, [Computer Software]

[5] Chafe, W. L.: Givenness, contrastiveness, definiteness, subjects, topics and point of view. In Charles N. Li, Subject and Topic. Academic Press: New York, 1976, pp. 27-55.

[6] Cruttenden, A.: The deaccenting of old information: A cognitive universal? In: G. Bernini & M.L. Schwarz (eds.) Pragmatic Organization of Discourse in the Languages of Europe. Gruyter, 2006, pp. 311-355.

[7] Fanselow, G. and D. Lenertová: Left Peripheral Focus. Mismatches between Syntax and Information Structure. Nat Lang Linguist Theory 29, 2011, pp. 169–209.

[8] Féry, C.: German Intonational Patterns, Niemeyer: Tübingen,1993.

[9] Féry, C. and F. Kügler: Pitch accent scaling on given, new and focused constituents in German. Journal of Phonetics 36, 2008, pp. 680–703.

[10] Féry, C. and S. Ishihara: How focus and givenness shape prosody. In M. Zimmermann, & C. Féry (Eds.), Information Structure. Theoretical, Typological, and Experimental Perspectives. Oup: Oxford, 2010, pp. 307–331.

[11] Götze, M., T. Weskott, C. Endriss, I. Fiedler, S Hinterwimmer, S. Petrova, A. Schwarz, S. Skopeteas and R. Stoel, Information Structure. In Dipper, S., Götze, M. & Skopeteas, S. (Eds.) Working Papers of the SFB632, (ISIS) 7, (Potsdam, Germany), 2007, pp. 147-187.

[12] Grice, M., S. Baumann and R. Benzmüller: German Intonation in Autosegmental-metrical Phonology. In Jun, Sun-Ah., Prosodic Typology: The Phonology of Intonation and Phrasing. OUP: Oxford, 2005, pp. 55-83.

[13] Grice, M., S. Baumann and N. Jagdfeld: Tonal association and derived nuclear accents: the case of downstepping contours in German. Lingua 119, 2009, pp. 881-905.

[14] Krifka, M.: Basic Notions of Information Structure. In Féry, C., Fanselow, G., Krifka, M. (eds.), Working Papers of the SFB632, Interdisciplinary Studies on Information Structure (ISIS) 6. (Potsdam, Germany), 2007, pp. 13–56.

[15] Kruijff, G.-J. M.: A Categorial Modal Architecture of Informativity: Dependency Grammar Logic & Information Structure. Ph.D. thesis, Charles University, Prague, Czech Republic. 2001.

[16] Kügler, F., C. Féry and R. van de Vijver: Pitch accent realization in German. Proceedings of the 15th ICPhS. (Barcelona, Spain), 2003, pp. 1261-1264.

[17] Kügler, F.: The role of duration as a phonetic correlate of focus. In Barbosa, P. A., Madureira, S., Reis, C. (eds.), Proceedings of the Speech Prosody 2008 Conference. Campinas, (Brazil), 2008, pp. 591–594.

[18] Kügler, F., Smolibocki, B., Stede, M.: Evaluation of Information Structure in Speech Synthesis: The Case of Product Recommender Systems. ITG Conference on Speech Communication: IEEE, 2012, pp. 127-131.

[19] Kügler, F. and Féry, C. (submitted): Post-focal downstep in German.

[20] Ladd, D.R.: Intonational Phonology. CUP, 1996.

[21] OpenCCG natural language processing library: http://openccg.sourceforge.net/

[22] Reiter, E. and R. Dale: Building Natural Language Generation Systems. Cambridge University Press. 2000. Reissued in paperback in 2006.

[23] Repp, S.: Defining 'contrast' as an information-structural notion in grammar. Lingua 120 (6), 2010, pp. 1333-1345.

[24] Romanelli, M.: Modeling givenness and contrast in MARY (Modular Architecture for Research on speech sYnthesis). Ms. Saarland University, 2003, http://www.dfki.de/~romanell/maxS2.pdf (accessed on 31. January 2013).

[25] Schröder, M. and J. Trouvain: The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. International Journal of Speech Technology 6, 2003, pp. 365–377.

[26] Stede, M.: Polibox: Generating desciptions, comparisons, and recommendations from a database. In Proceedings of the 19th Int'l Conference on Computational Linguistics (Coling), (Taipei), 2002.

[27] Steedman, M. and J. Baldridge: Combinatory Categorial Grammar. In: R. Borsley and K. Borjars (eds.) Non-Transformational Syntax. Blackwell: Oxford, 2011, pp. 181-224.

[28] Tokuda, K., H. Zen and A.W. Black: An HMM-based speech synthesis system applied to English, Proc. of 2002 IEEE SSW, Sep. 2002.

[29] Uhmann, S.: Fokusphonologie. Eine Analyse deutscher Intonationskonturen im Rahmen der nicht-linearen Phonologie. Tübingen: Niemeyer, 1991.

[30] Xu, Y.: Effects of tone and focus on the formation and alignment of F0 contours. Journal of Phonetics 27, 1999, pp. 55-105.

Appendix

Context/Target sentence as answer to context question:

1.1 Hat sich Luna ein iPhone gekauft? Martin hat sich ein iPhone gekauft.

1.2 Hat sich Thomas Bücher gekauft? Lisa hat sich Bücher gekauft.

1.3 Hat sich Mona Bilder gekauft? Martin hat sich Bilder gekauft.

1.4 Hat sich Markus Hosen gekauft? Martin hat sich Hosen gekauft.

2.1 Kaufte Mona ein Handy im Internet? Martin kaufte ein Handy im Internet.

2.2 Funktioniert das Nokia Lumia im Netz von T-Mobile? Das iPhone funktioniert im Netz von T-Mobile.

2.3 Bemalt Markus die Wände in der Wohnung? Lisa bemalt die Wände in der Wohnung.

2.4 Sammelt Dagmar die Samen der Blumen? Lisa sammelt die Samen der Blumen.