Integrating Contrast in a Framework for Predicting Prosody

Pepi Stavropoulou, Dimitris Spiliotopoulos, and Georgios Kouroupetroglou

Department of Informatics and Telecommunications,
National and Kapodistrian University of Athens,
Panepistimiopolis, Ilisia, GR-15784, Athens, Greece
{pepis,dspiliot,koupe}@di.uoa.gr

Abstract. Information Structure (IS) is known to bear a significant effect on Prosody, making the identification of this effect crucial for improving the quality of synthetic speech. Recent theories identify contrast as a central IS element affecting accentuation. This paper presents the results of two experiments aiming to investigate the function of the different levels of contrast within the topic and focus of the utterance, and their effect on the prosody of Greek. Analysis showed that distinguishing between at least two contrast types is important for determining the appropriate accent type, and, therefore, such a distinction should be included in a description of the IS – Prosody interaction. For this description to be useful for practical applications, a framework is required that makes this information accessible to the speech synthesizer. This work reports on such a language-independent framework integration of all identified grammatical and syntactic prerequisites for creating a linguistically enriched input for speech synthesis.

Keywords: Information Structure, Contrast, Prosody Prediction, Speech Synthesis, Annotation Framework.

1 Introduction

It is generally acknowledged that there is a significant interaction between Information Structure (IS) and Prosody. Identifying this interaction is, therefore, very important in the case of practical applications such as speech synthesizers, whereas the quality of the prosody of the utterance greatly determines the overall quality, naturalness and legibility, of the synthetic speech. In addition to the fundamental information-structural partition of the utterance into topic and focus (or theme and rheme, or topic and comment etc. depending on the approach) recent theories [3, 6, 13] identify contrast as a significant IS element claimed to affect accentuation. Furthermore, several researchers [5, 8] propose the existence of different types – or alternatively a hierarchy – of contrast, based on evidence from various languages that grammatically encode different levels of this contrast hierarchy. This paper presents an empirical study of the effect of the various levels of contrast on the prosody of

A. Gelbukh (Ed.): CICLing 2010, LNCS 6008, pp. 748–758, 2010.

Modern Greek and further discusses the integration of this meta-information for creating a linguistically enriched text description for prosody prediction in speech synthesis into an appropriate framework.

1.1 Theoretical Background

Two-dimensional views of Information Structure identify: (i) a high level partition of the utterance into complementary parts, such as topic and focus, and (ii) a lower level mechanism that functions both within the topic and the focus part of the utterance and is associated with some notion of contrast [6, 13, 16] or givenness [3]. Contrast, in this case, is related to the possibility of different, alternative referents made available by the context, and is marked by a pitch accent as opposed to background material, which remains unmarked. Sentence (1) illustrates this two-level distinction. Prosodically prominent words are capitalized.

(1) What did the tourists want?
The British tourist wanted to rent the blue car. [The ITALIAN_{C} tourist]$_{TOPIC}$ [wanted to rent the RED_{C} car]$_{FOCUS}$.

In this more semantically-oriented, quantification-based view of contrast, every focus is contrastive as it triggers the presupposition of a set of alternatives to the focused element. Even in cases of broad focus, one may argue that it is one state of affairs that is contrasted with another [4, 9]. Some researchers, however, combining a more pragmatic or "informational" approach, argue for the existence of different types of contrast, each one of which may be differently encoded in the structure of the language, bearing distinct prosodic, morphological or syntactic correlates. [8] proposes the following criteria for the definition of a hierarchy of contrast (from weaker to stronger): mere *highlighting* through accentuation → existence of a *dominant contrast*, dividing the utterance into a focus and background part → existence of an *open set of alternatives* → existence of a *limited closed set of alternatives* → *explicit mentioning of alternatives* in the context (i.e. existence of a salient directly accessible set). In addition to these criteria, *correction* has been proposed as a special case of contrast that has distinct prosodic markers [5, 6]. It is actually the case that – in some languages at least – only correction as opposed to other sub-notions of contrast is expressed differently.

The different levels of this contrast hierarchy are associated with different types of topics or foci as shown in Table 1. Accordingly, the primary descriptive goal of the study presented here is to examine the prosodic correlates of the different types of topics and foci, ultimately identifying the levels of contrast that are encoded in the prosody of Modern Greek. Furthermore, this study aims to assess the range of interaction between contrast and the topic–focus partition, in order to identify the type of information that should be integrated in a framework for predicting prosody. That is, if the notion of contrast alone is enough to determine accentuation, then it should be formally represented as an autonomous IS feature and there would be no need to resolve to the identification of different types of foci or topics.

Table 1. Association of contrast with different types of topic and focus

	High-lighting	**Dominant Contrast – Open Set of Alternatives**	**Salient Closed Set of Alternatives**	**Correction**
All New / Topic-less Utterances, Broad Information Focus	+	-	-	-
Narrow Information Focus	+	+	-	-
Simple Topic	(+)	+	-	-
Contrastive Focus	+	+	+	-
Contrastive Topic	+	+	+	-
Corrective Focus	+	+	+	+
Corrective Topic	+	+	+	+

2 Experimental Setup

To address these issues two pilot experiments were carried out, the first one investigating the effect that different types of topics have on prosody, and the second one investigating the effect of different types of foci.

2.1 Experiment A - Topics

Three types of topics were tested: simple, contrastive and corrective topics. Sentences (2), (3) and (4) are examples of each type respectively.

(2) What did the Italian tourist want?
[The Italian tourist]$_{ST}$ wanted to rent a car [Simple Topic]

(3) What did the tourists want?
The British tourist wanted to rent a room,
[the ITALIAN tourist]$_{ConT}$ wanted to rent a car [Contrastive Topic]

(4) What did the British tourist want?
[The ITALIAN tourist]$_{CorT}$ wanted to rent a car [Corrective Topic]

All types were compared against all new / topic-less utterances as well. Therefore four pragmatic conditions in total were examined. Test material consisted of 7 utterances per condition. Each utterance was produced twice, once following a narration and once following a Q/A disambiguating context. All utterances were produced by 9 speakers of Athenian Greek resulting in 504 (4x7x2x9) tokens in total. Speakers read the material in random order. Topics were sentence-initial, one and two content-word phrases. To avoid topic accommodation in all new sentences, a generic

version of the utterances was used for the no topic condition; that is an indefinite noun phrase was used instead of a definite one, as definitiveness is often assumed to signal knowledge already present in the hearer's knowledge store.

The four pragmatic conditions were compared on the basis of both phonological and phonetic criteria. In the first case, utterances were annotated for pitch accent type based on the GRToBI annotation scheme [1]. In the second case, measurements were taken of mean F0 (vowel), vowel duration and mean intensity (vowel). Statistical significance was tested using chi-square tests and ANOVAs for phonological and phonetic values respectively.

2.2 Experiment B - Foci

As in the case of topics, 4 pragmatic conditions were tested for focus as well: all new sentences, (narrow) information focus, contrastive focus and corrective focus. Sentences (5)-(8) are examples of the focus types examined.

(5) (What's going on?)
The mailman is looking for HELEN [Broad Focus-All New]
(6) Who is the mailman looking for?
The mailman is looking for $\text{HELEN}_{\text{InfF}}$ [Information Focus]
(7) Who is the mailman looking for? Michael or Helen?
The mailman is looking for $\text{HELEN}_{\text{ConF}}$ [Contrastive Focus]
(8) The mailman is looking for Michael
(No), the mailman is looking for $\text{HELEN}_{\text{CorF}}$ [Corrective Focus]

Seven sentences per condition were embedded in disambiguating contexts to be produced in random order by 5 speakers of Athenian Greek, resulting in a total of 140 (4x7x5) tokens. Focus phrases were always sentence final. Materials were annotated for pitch accent type, and measurements of mean F0 (vowel), vowel duration and mean intensity (vowel) were taken and subjected to analysis of variance. Chi-square tests were used to calculate the effect on pitch accent.

3 Results

The L+H* pitch accent was the predominant choice for both corrective topic and corrective focus. The L* and H* were the accents most commonly used for the remaining types of topic and focus respectively. Figure 1 shows the distribution of nuclear pitch accents over the four pragmatic conditions examined in experiments A and B. Accent distribution proved to be statistically significant for all speakers in the case of topics ($p<0.0005$) and for all speakers (ranging from $p<0.001$ to $p<0.008$ depending on speaker) but one ($p<0.634$) in the case of foci. That one speaker resorted to an emphatic rendition for all utterance types.

Moreover, corrective topics were uttered with increased intensity, duration and F0. All dependent variables showed statistically significant effect ([$F(3)=47.825$, $p<0.0005$], [$F(3)=23.505$, $p<0.0005$], [$F(3)=417.944$, $p<0.0005$] for mean intensity, duration, and mean F0 respectively). Post hoc Turkey tests revealed that only

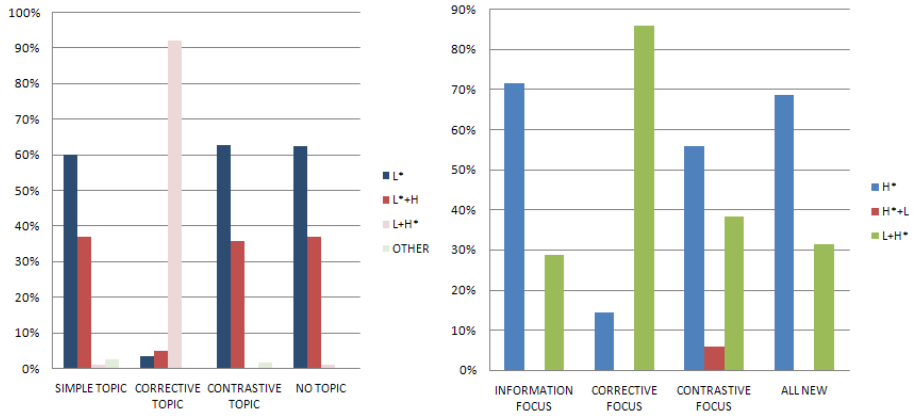

Fig. 1. Distribution of pitch accents over topic and focus types

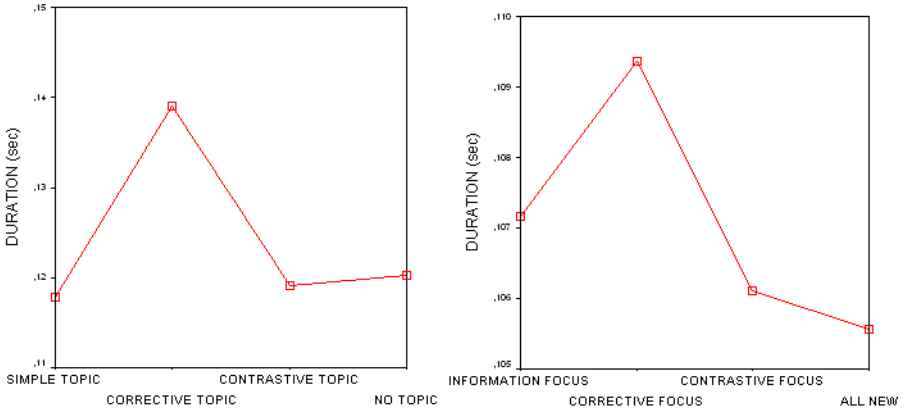

Fig. 2. Mean Vowel Duration for different topic and focus types

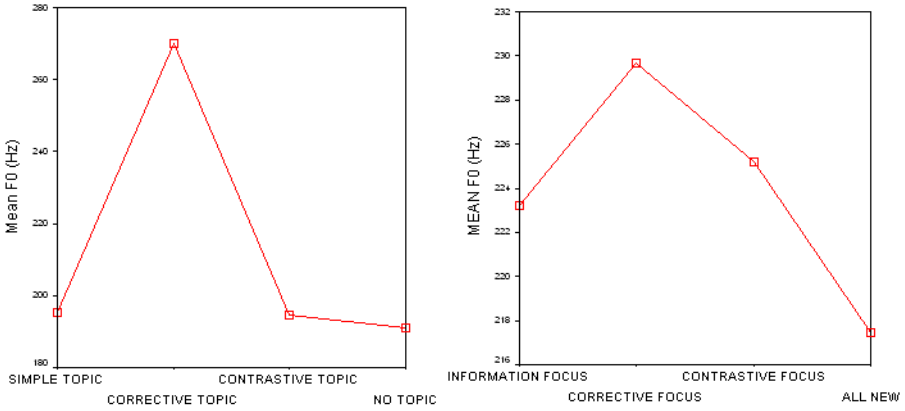

Fig. 3. Mean F0 (vowel) for different topic and focus types

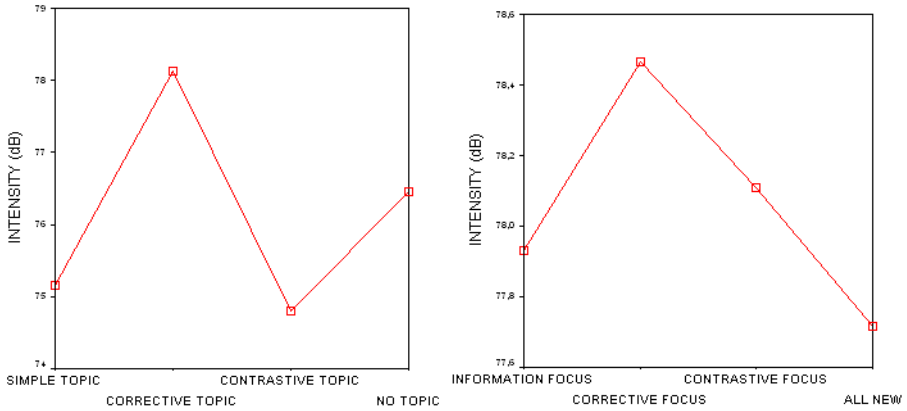

Fig. 4. Mean intensity (vowel) for different topic and focus types

corrective topics significantly differed in pair-wise comparisons, except for the case of intensity, whereas topic-less phrases also differed. In the case of focus, on the other hand, only F0 differed with marginal statistical significance [$F(3)=1756$, $p<0.018$]. Figures 2-4 summarize the results.

4 Discussion

The results of the experiments presented here show that only corrective topics and foci are clearly and consistently distinguished from the other three conditions on the basis of both phonological (L+H* pitch accent) and phonetic (increased intensity, duration, F0 for topics, and F0 for foci) properties. Therefore, Greek only seems to mark correction – with regards to intonation at least – as opposed to other levels of contrast. This does not come as a surprise, as – from an "informational" point of view [15] – correction is the most cognitively loaded procedure, involving subtraction as well as addition of information to the hearer's knowledge store. Similar behavior has been observed in several languages, whereas only corrective focus – as opposed to other types of foci – has distinct phonological correlates, and is therefore structurally contrastive [5]. Moreover, correction is associated with the feature of exhaustivity [7] (i.e. the identification of a unique and maximal subset from the set of alternatives, for which subset only, the predicate phrase actually holds), which in turn has been associated with identificational focus [17]. Identificational focus is an additionally marked case of focus as, on top of being contrastive, is exhaustive as well.

Furthermore, analysis showed that the same nuclear pitch accent (NPA) was used for corrective topic as well as corrective focus, suggesting that the marked effect of correction is independent of the topic-focus articulation, at least with regards to the type of NPA employed. The significant increase in duration and intensity that was observed for corrective topics only, could be explained on the basis of their sentence-initial position (cf. [10]) rather than as being a reflection of topichood. In short, one

could argue that it is not corrective topic or focus per se that is expressed differently, but that the difference is due to the low-level contrast feature that functions within both topic and focus and that topichood or focusing do not determine accent type in contrast to what has been suggested in the literature [13, 14]. The above argument is corroborated by the fact that previous work [2] has shown that, for Greek, the tonal pattern for topic in declaratives is the same as the tonal pattern for focus in interrogatives and vice versa, suggesting that it is the boundary tone that "selects" NPA type, ultimately associating the latter to the discourse role of the former, further disassociating NPA type from topichood or focusing. Similarly, our analysis showed that the contour used for all new phrases was the same for simple and contrastive topics, further supporting the claim that it is not the topic-focus distinction that is conveyed through pitch accent type. As a result, the L* and H* accents that were the predominant choice for the remaining types of topic and focus in our corpus cannot be considered as a constant marker of topichood or focusing.

Even though only correction, compared to other types of contrast, seems to be able to determine the NPA type, identifying what is contrastive in the broad semantically oriented view of contrast, is still necessary in order to define the location of the nuclear pitch accent. That becomes clear in the case of deaccenting, whereas the word which distinguishes the focused element from other alternatives carries the Nuclear Pitch Accent causing all following words to surface de-accented. In some models of Information Structure [3, 11, 12], this function of contrast is ascribed to the function of givenness, whereas a given element is informally defined as an element that has been previously mentioned or can be entailed from another previously mentioned constituent. The prosodic effect is the same, whether it is alternative entities that are distinguished or new vs. given elements. In a similar vein, [12] proposes the postulation of two different features, G and F, in the syntactic representation of the utterance, which correspond to givenness and contrast respectively. It is claimed that the combination of these two features can adequately describe different, structurally motivated types of topics or foci.

In the following section, we will present a markup framework for prosody prediction, whereas pragmatic contrast – i.e. correction in the case of Greek – is represented as an autonomous feature and semantic contrast in the broad sense is conveyed through the given-new distinction. It should be noted that while correction seems to be the minimum pre-requisite for contrastive marking in Greek, other languages may still be "structurally sensitive" to other levels lower in the contrast hierarchy.

5 Integration to an Annotation Framework

Speech synthesizers traditionally perform a part-of-speech analysis and build the syntactic tree of the text in order to assign prosody [18]. General purpose Text-to-Speech (TtS) systems use certain language processing subsystems, such as sentence segmentation and part-of-speech tagging, for the analysis of the written text input.

Depending on the actual system, such analysis may suffer from inherent statistical error accuracy that may be due to the design and implementation of the respective modules or language ambiguity. However, TtS systems may employ language analysis modules that are designed for high accuracy in specific thematic domains for which they seem to perform adequately. The respective accuracy when used for generic or other thematic domains may fall under unacceptable levels. Additionally, the language processing modules embedded in TtS systems are not usually designed to identify and extract higher-level linguistic information, such as semantic or pragmatic factors, that may be used to aid speech synthesis.

Previous works that have explored prosody and speech synthesis show that linguistically enriched annotated text input to a speech synthesizer can lead to improved naturalness of speech output [19, 20]. Generation of tones and prosodic phrasing from high level linguistic input produces better prosody than plain texts do [21]. When such input can be provided, the language processing from the TtS system can be superseded. In this respect, integrating contrast into a framework for language analysis and semantic annotation is important in order to produce an enriched text description as input for speech synthesizers. Text annotation is a procedure where certain meta-information gets identified and associated with the entities in a text corpus. Such information is commonly used in computational linguistics for language analysis, speech processing, natural language processing, speech synthesis, and other areas. The type of information that is analyzed and associated to text units may span the linguistic analysis tree (grammatical, syntactic, morphological, semantic, pragmatic, phonological, phonetic), as well as include any other description that may be of use.

Existing frameworks included the feature and annotation of *contrast* as a process rule [22]. The other features that are currently used for determining the intonational focus prominence include *newness (new or old information)*, *explicit emphasis*, *first or second argument to verb, proper- or common-noun*. Extending that description, based on the aforementioned results, contrast may be included as two distinct features, each providing a more accurate respective prosodic manipulation. Consider the following sentences taken from [22]:

(9) This exhibit was made$_{\text{New}}$ in Beotea$_{\text{New}}$.
[It was found$_{\text{New}}$ in Beotea$_{\text{Giv}}$ but it was made$_{\text{Giv}}$ in Athens$_{\text{New}}$]$_{\text{CONSTRAST}}$

The analysis of the corrective vs informational contrast dictates that the contour of sentence 9 should be treated differently to the prototypical contour of the corresponding all-new sentence. Focus prominence and pitch accent prediction shifts from the proper-noun “Beotea” to the verb “found” in the first clause, and proper-noun “Athens” receives special emphasis when corrective contrast is introduced. Providing a distinction between corrective and all other types of contrast, the annotation of this feature can result in proper prosody prediction of those instances. Informational contrast can be described by the newness factor while corrective should be a distinct feature. For Greek, as a generalisation rule, contrast is used for correction while all other instances are described by association with new/given information feature.

```
<utterance>
<relation name="Word" structure-type="list">
<wordlist>
<w id="w01">It</w>
<w id="w02">was</w>
<w id="w03">found</w>
<w id="w04">in</w>
<w id="w05">Beotea</w>
<w id="w06">but</w>
<w id="w07">it</w>
<w id="w08">was</w>
<w id="w09">made</w>
<w id="w10">in</w>
<w id="w11" punct=".">Athens</w>
</wordlist>
</relation>
<relation name="Group" structure-type="list">
</relation>
<relation name="Syntax" structure-type="tree">
<elem phrase-type="S">
<elem phrase-type="prosody" event="contrast">
<elem lex-cat="PRONOUN" href="#w01"/>
<elem lex-cat="AUX" href="#w02"/>
<elem phrase-type="prosody" newness="true" class="mid-emphasis-verb">
<elem lex-cat="VERB" href="#w03"/>
</elem>
<elem lex-cat="PREPOS" href="#w04"/>
<elem phrase-type="prosody" newness="false" arg="arg2" class="proper-
noun">
<elem lex-cat="NOUN" href="#w05"/>
</elem>
<elem phrase-type="prosody" class="mid-emphasis-conj">
<elem lex-cat="CONJNCT" href="#w06"/>
</elem>
<elem lex-cat="PRONOUN" href="#w07"/>
<elem lex-cat="AUX" href="#w08"/>
<elem phrase-type="prosody" newness="false", class="mid-emphasis-verb">
<elem lex-cat="VERB" href="#w09"/>
</elem>
<elem lex-cat="PREPOS" href="#w10"/>
<elem phrase-type="prosody" newness="true" arg="arg2" class="proper-
noun">
<elem lex-cat="NOUN" href="#w11"/>
</elem>
</elem>
</elem>
</relation>
</utterance>
```

Fig. 5. The XML description

Figure 5 shows the XML output for the sentence *"It was found in Beotea but it was made in Athens"* as annotated within the framework and exported to XML. First part is a wordlist of all tokens (words) and punctuation values (<*wordlist*>), followed by the syntax tree, prosodic features, and other high-level information (<*relation*>). This is the input for the speech synthesizer that contains meta-information about how

contrast is assigned as a property of the whole phrase and is subsequently associated with the particular new word within the sentence.

6 Conclusion

The empirical evidence presented in this paper favors the postulation of two different types of contrast as predictors for prosody generation. The two types are associated with a semantic view of contrast, whereas all utterances are in a broad sense contrastive, and a pragmatic one respectively. The latter is a feature of certain utterances only that fulfill specific conditions. The minimum conditions required are subject to typological parameterization, as different languages may express different levels of pragmatic contrast. Greek in particular seems to be sensitive to correction, the level with the highest cognitive load. In the text processing framework described here semantic contrast is accommodated through the given-new distinction and pragmatic contrast is represented as an additional autonomous feature.

Acknowledgments

The work described in this paper has been funded by the Special Account for Research Grants of the National and Kapodistrian University of Athens under the KAPODISTRIAS program.

References

1. Arvaniti, A., Baltazani, M.: Intonational Analysis and Prosodic Annotation of Greek Spoken Corpora. In: Jun, S.-A. (ed.) Prosodic Typology: The Phonology of Intonation and Phrasing, pp. 84–117. Oxford University Press, Oxford (2005)
2. Baltazani, M., Jun, S.-A.: Focus and topic intonation in Greek. In: Proceedings of the 14th International Congress of Phonetic Sciences, vol. 2, pp. 1305–1308 (1999)
3. Büring, D.: Semantics, Intonation and Information Structure. In: Ramchad, G., Reiss, C. (eds.) The Oxford Handbook of Linguistic Interfaces. Oxford University Press, Oxford (2007)
4. Dretske, F.J.: Contrastive statements. The Philosophical Review 81, 411–437 (1972)
5. Gussenhoven, C.: Types of Focus in English. In: Lee, C., Gordon, M., Büring, D. (eds.) Topic and Focus: Cross-linguistic Perspectives on Meaning and Intonation, pp. 83–100. Springer, Heidelberg (2007)
6. Krifka, M.: Basic notions of information structure. In: Fery, C., Krifka, M. (eds.) Interdisciplinary Studies of Information Structure, Potsdam, vol. 6 (2007)
7. Van Leusen, N., Kalman, L.: The Interpretation of Free Focus. In: ILLC Computational Linguistics (1993)
8. Molnár, V.: Contrast from a contrastive perspective. In: Kruiff- Korbayová, I., Steedman, M. (eds.) ESSLLI 2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics (2001)
9. Rooth, M.: A Theory of Focus Interpretation. Natural Language Semantics 1, 75–116 (2001)

10. Rump, H., Collier, R.: Focus Conditions and the prominence of pitch-accented syllables. Language & Speech 39, 1–17 (1996)
11. Schwarzschild, R.: GIVENness, AvoidF and Other Constraints on the placement of Accent. Natural Language Semantics 7(2), 141–177 (1999)
12. Selkirk, E.: Contrastive Focus, Givenness and the Unmarked Status of "Discourse-New". In: Féry, C., Fanselow, G., Krifka, M. (eds.) Working Papers of the SFB632, Interdisciplinary Studies on Information Structure (ISIS), vol. 6, pp. 125–146. Universitätsverlag Potsdam, Potsdam (2007)
13. Steedman, M.: Information structure and the syntax-phonology interface. Linguistic Inquiry 31, 649–689 (2000)
14. Steedman, M.: Information-Structural Semantics of English Intonation. In: Gordon, M., Büring, D., Lee, C. (eds.) LSA Summer Institute Workshop on Topic and Focus, Santa Barbara, pp. 245–264. Kluwer Academic, Dordrecht (2002)
15. Vallduví, E.: The Informational Component. Garland Publishers, New York (1992)
16. Vallduví, E., Vilkuna, M.: On Rheme and Kontrast. In: Culicover, P., Wagner, M. (eds.) Givenness and Locality. The Limits of Syntax, pp. 79–108. Academic Press, San Diego (1998)
17. Kiss, K.E.: Identificational focus versus information focus. Language 74, 245–273 (1998)
18. Taylor, P., Black, A., Caley, R.: The architecture of the festival speech synthesis system. In: Proc. 3rd ESCA Workshop on Speech Synthesis, Australia, pp. 147–151 (1998)
19. Pan, S., McKeown, K., Hirschberg, J.: Exploring features from natural language generation for prosody modeling. Computer Speech and Language 16, 457–490 (2002)
20. Xydas, G., Spiliotopoulos, D., Kouroupetroglou, G.: Modeling Improved Prosody Generation from High-Level Linguistically Annotated Corpora. IEICE Trans. of Inf. and Syst., Special Section on Corpus-Based Speech Technologies 88(3), 510–518 (2005)
21. Black, A., Taylor, P.: Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input. In: Proc. 3rd Int. Conf. on Spoken Language Processing, Yokohama, Japan, pp. 715–718 (1994)
22. Spiliotopoulos, D., Petasis, G., Kouroupetroglou, G.: A Framework for Language-Independent Analysis and Prosodic Feature Annotation of Text Corpora. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2008. LNCS (LNAI), vol. 5246, pp. 517–524. Springer, Heidelberg (2008)