WHERE GREEK TEXT TO SPEECH FAILS
REQUIREMENTS FOR SPEECH SYNTHESIS IN SPOKEN DIALOGUE SYSTEMS

Pepi Stavropoulou[1,2], Dimitris Spiliotopoulos[2] and Georgios Kouroupetroglou[2]
[1]*Department of Philology*
University of Ioannina
Ioannina, Greece
[2]*Speech and Accessibility Lab.*
Department of Informatics and Telecommunications
University of Athens
Athens, Greece
{pepis,dspiliot,koupe}@di.uoa.gr

Abstract

This paper sets out the key requirements for effective use of Text to Speech (TtS) synthesis in automated spoken dialogue systems. It identifies basic shortcomings of current TtS systems in human-machine, task oriented Greek dialogues. It further verifies and completes phonological descriptions of Greek prosody with regards to the specific genre, focusing particularly on list structures. Finally, it takes a first step towards proposing a pragmatically motivated annotation schema that could help achieve a more accurate prosody specification and consequently a more natural TtS rendition.

Keywords: TtS, list intonation, prosody, Concept to Speech synthesis

1. Introduction

Despite the high cost of voice talents, studio time, and the occasional rigidity of pre-recorded, pre-determined scripts, the vast majority of commercial automated spoken dialogue systems resort to pre-recorded acted prompts, instead of using Text to Speech (TtS), as a result of inadequate performance of current TtS systems. An important source of errors is inappropriate intonation. Most generic TtS systems are trained on neutral, read speech databases, which both differ in style and lack pragmatic events often occurring in the dialogue (Syrdal and Kim, 2008). Furthermore, TtS systems (Fellbaum and Kouroupetroglou, 2008) typically do not take various important aspects of context into account, which have been shown to greatly affect prosody (Büring, 2007;2010; Baltazani, 2006 among many others).

On the other hand, automated dialogue systems (cf. Figure 1) could readily provide the TtS module with a much richer, context-aware input than plain text to be rendered to speech (Xydas et al., 2003a). As part of a Concept to Speech (CtS) synthesis process (Young and Fallside, 1979; Taylor, 2000; McKeown and Pan, 2000; Xydas et al., 2003b), the input to the synthesizer could also include important, error free linguistic information regarding the structure and the context of the utterance, effectively guiding its prosodic rendition. This information could range from e.g. part of speech information and syntax role to pragmatic events such as dialogue acts and focusing (Xydas et al., 2004; 2005; Spiliotopoulos et al., 2008).

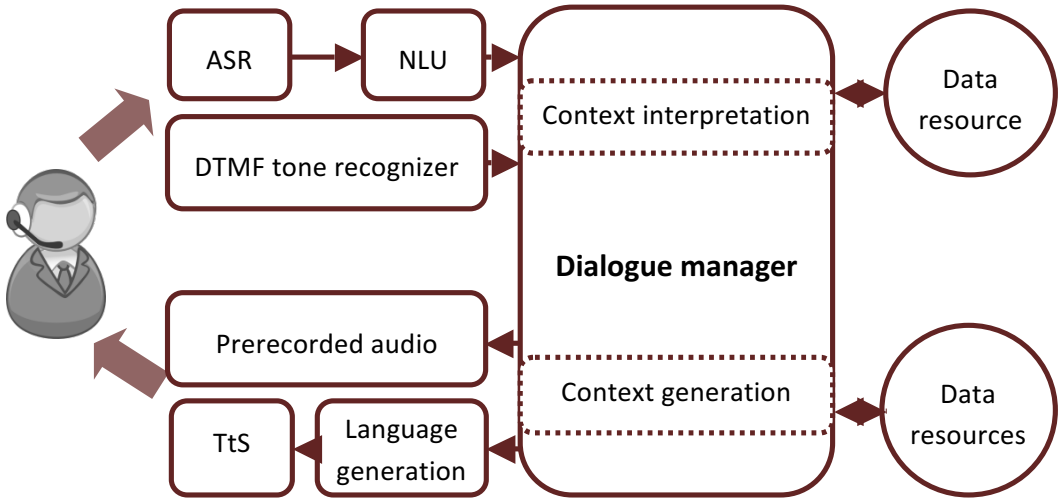

Figure 1 An automated dialogue system's main components (Automated Speech Recognition, Natural Language Understanding, Dialogue Manager, Language Generation and Text to Speech). Figure adapted from Stavropoulou et al. (2010).

This study takes a closer look into the exact requirements posed by the use of TtS in spoken dialogue systems. More specifically, it aims to address the following questions:

a) Which linguistic structures are most commonly used in automated dialogue systems in human computer interaction? How are these structures prosodically realized?
b) To what extent are these structures already effectively handled by current TtS synthesizers?

Answering the above questions can guide the specifications for both the extra linguistic information that is required as additional input to the synthesizer, as well as the content of the corpora used for training and developing TtS systems; ultimately this would lead to more natural output and effective use of TtS modules in automated spoken dialogue systems.

Accordingly, this paper presents the results of a linguistic analysis of automated dialogue system prompts focusing on the sentence type. Each type is mapped to respective prosodic realizations, and three state of the art Greek TtS synthesizers are evaluated with regards to each sentence type.

In the following sections we first outline the methodology followed (Section 2). Distribution analysis of sentence types in the particular genre is presented in Section 3. The results of the prosodic analysis of the most frequently occurring sentence types are presented in Section 4. Particular emphasis is placed on the analysis of list structures in Greek polars, which to our knowledge have not been systematically analyzed before. Section 5 presents the results of TtS system evaluation. Final section elaborates on key findings and presents a preliminary version of a pragmatic annotation schema that could enable effective data driven or rule based prosody prediction.

2. Methodology

2.1 System prompt analysis

We examined the system prompts (i.e. system utterances) of six typical commercial Greek automated dialogue systems, in order to identify the most commonly used sentence types and dialogue acts[1]. The systems were task oriented, ranging from ticketing services to banking and customer care services, and were both Directed Dialogue and Natural Language Understanding (NLU) systems (cf. Table 1). Task oriented dialogue systems were deemed most appropriate for our analysis, as they a) constitute the predominant type of dialogue systems in industry and b) the

[1] Analysis of dialogue acts is beyond the scope of this paper and will therefore not be attempted; we will present results on sentence type alone.

constrained structure of the task is optimal for unambiguous identification of context and information structure.

Table 1 Distribution of prompts per application and system type

Application	Application Domain	System Type	No of Prompts
A	Ticketing	Directed Dialogue	103
B	Banking Info / Transactions	Directed Dialogue	99
C	Info on mobile telephony products	Directed Dialogue	124
D	Mobile Telephony Shop Representatives Customer Care	NLU	253
E	Mobile Telephony Customer Care	NLU	433
F	Landline and Internet Customer Care	NLU	237

Both directed dialogue and NLU systems were included, to ensure that potential differences in the prompts of each system type are accounted for. Typically, directed dialogue systems use more directive, rigid, menu-like prompts, while NLU systems make use of open-ended questions (e.g. "How may I help you?") and more natural, "human-like" prompts as well.

Table 2 Examples of sentence types

Sentence Type	Examples
Declaratives – Plain	Η αίτησή σας βρίσκεται στο στάδιο ελέγχου. Your application is being processed.
Declaratives – Lists	Μπορείτε να ζητήσετε γενικές πληροφορίες, στοιχεία λογαριασμού ή υπόλοιπο χρήσης. You can ask for general information, billing information or free credit.
Declaratives - Negation	Ο αριθμός αυτός δεν υπάρχει στη βάση. The phone number is not in our database.
Polars - Plain	Ενδιαφέρεστε για θέματα λογαριασμών; Are you interesting in billing information?
Polars – Lists (exclusive OR)	Πρόκειται για συνδρομή συμβολαίου, ασύρματο ίντερνετ ή καρτοκινητό; Are you interested in a postpaid contract, prepaid or mobile internet?
Wh-Questions	Πού θέλετε να ταξιδέψετε; Where would you like to travel?
Imperatives – Plain	Προχωρήστε σε έκδοση χειρόγραφων αποδείξεων. Proceed to issue handwritten receipts.
Imperatives – Lists	Πείτε το όνομα του πλοίου, το όνομα του λιμανιού ή το δρομολόγιο που σας ενδιαφέρει. Tell me the name of the ship, the name of the port or the route you are interested in.

The following eight sentence types were used for the analysis: 1) Declaratives-Plain, 2) Declaratives-Lists, 3) Declaratives-Negation, 4) Polars-Plain, 5) Polars-Lists, 6) Wh-Questions, 7) Imperatives-Plain, 8) Imperatives-Lists. Examples for each type are presented in Table 2.

The non-standard types involving lists, namely "Declarative - Lists", "Polars - Lists" and "Imperatives - Lists", were included as distinct sentence types, because of the wide distribution and importance of lists in the particular genre. For the purposes of this study, lists are defined as sequences of two or more constituents of the same type, the last of which is typically introduced with the operator "or".

2.2 Prosodic analysis

Next we analyzed how system prompts corresponding to the identified linguistic structures were prosodically produced by 4 experienced female actors. The analysis was based on GRToBI (Arvaniti & Baltazani, 2005), and had a twofold aim:

a) To validate existing descriptions of Greek phonology with respect to this specific genre and style.
b) To identify and phonologically describe structures that have not yet been extensively described.

Prompts had been recorded under the supervision of an expert linguist specializing in spoken dialogue interfaces and actor coaching. Actors were provided with written scripts/scenarios, where each prompt was placed in the intended context. Recordings were conducted in a sound proof booth. Audio signal was digitized at 44100 Hz using 16-bit samples. Waveform analysis was performed in Praat (Boersma and Weenink, 2005). 382 utterances were analyzed in total. Their distribution per sentence type is shown in Table 3.

2.3 Evaluation

Last, we compared the acted prompts to the corresponding output of 3 state of the art Greek TtS systems. The TtS synthesizers evaluated were: Loquendo 7 (Artemis/Afroditi), Vocalizer 5 (Melina) and Acapela 9 (Dimitris). The comparison was made on the basis of 32 utterances (4 utterances per sentence type; 8 sentence types). The evaluation was performed by an expert linguist. TtS productions were compared to the corresponding acted prompts, and each production was categorized as acceptable or not acceptable. Care was taken to ensure that input from previous preprocessing stages (e.g. word pronunciation specification) was error free.

Table 3 Distribution of sentence types produced by each actor

	Actor A	Actor B	Actor C	Actor D	Total
Declaratives – Plain	7	9	17	23	56
Declaratives – Lists	3	2	1	17	23
Declaratives - Negation	8	10	8	16	42
Declaratives - Total	*18*	*21*	*26*	*56*	***121***
Polars - Plain	2	22	9	43	76
Polars - Lists	7	8	16	36	67
Polars-Total	*11*	*32*	*26*	*81*	***150***
Wh-Questions	13	6	5	10	**34**
Imperatives-Plain	10	10	16	29	65
Imperatives-Lists	7	0	6	6	19
Imperatives-Total	*17*	*10*	*22*	*35*	***84***
Utterances - Total	*57*	*67*	*78*	*180*	***382***

3. Distribution of sentence types

Figure 2 presents the distribution of sentence types for the six applications examined. As shown in figure 2, there was a high percentage of questions and imperatives (38% and 23% respectively). In NLU systems the percentage of questions was even higher amounting to 45%. There was also a high percentage of list structures (18%), which we would normally not expect to find in most genres. In fact, the distribution of sentence types poses certain requirements on the content of the speech databases used for developing the TtS system, and these requirements are typically not met in the read speech databases used for developing generic synthesizers. Finally, it should be noted that there were also cases of non default early focus position, which typically triggers the deaccenting of material following the focused word. Such structures require that the TtS systems utilize contextual information in order to be able to produce appropriate, context sensitive intonation contours.

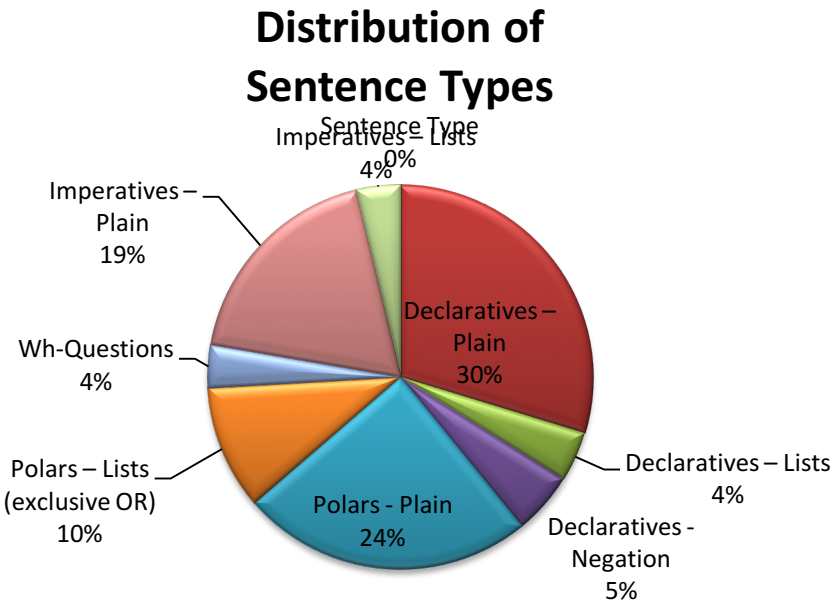

Figure 2 Distribution of sentence types corresponding to automated dialogue system prompts.

4. Prosodic Analysis Results

The prosodic analysis results, in most cases, confirm previous descriptions of the phonological melody of declaratives, negation, polar questions and wh-questions (Arvaniti and Baltazani, 2005; Arvaniti, 2007). Furthermore, our analysis revealed consistency in the realization of list structures in Greek, which – to our knowledge – have not yet been systematically described.

4.1 Declaratives and Imperatives

Declaratives and imperatives were associated with similar contours and are thus presented together. More specifically, both sentence types were produced with a final L-L% phrase/boundary tone combination. Nuclear pitch accents varied between H*, H*+L and L+H*, with H* being the most common one (45,57% and 52.63% for declaratives and imperatives respectively). Whilst no obvious structural differences appeared to affect the choice between the H* and H*+L accents, the L+H* accent was consistently linked to non final, early focus position. The above are in accordance with the observations in Arvaniti and Baltazani (2005), where the H*+L accent is associated with paralinguistic events and the L+H* with narrow focus, as well as Stavropoulou (2013), where sentence initial focus is shown to consistently carry a L+H* accent.

4.1.1 Lists in Declaratives and Imperatives

As expected, list structures posed strong constraints on the prosodic phrasing of the utterance. In particular, each list item triggered an intonational phrase boundary to its right, and was typically produced within its own intonational phrase. The initial list item typically but not always introduced a boundary to its left, while the operator "or" was also in some cases produced as a separate intonational phrase. The latter constitutes an arguably clear indication of an over-enunciating, emphatic speech style, which is characteristic of the specific genre.

Overall, there were two main strategies for the realization of lists in declaratives and imperatives (cf. figures 3 and 4):

a) All non final elements were realized with a high boundary (L-H%, H-H%, L-!H%), while the last element was realized with a low boundary (L-L%). This strategy was the most frequent one (74%).
b) Both final and non final elements ended in a low boundary (L-L%).
c)

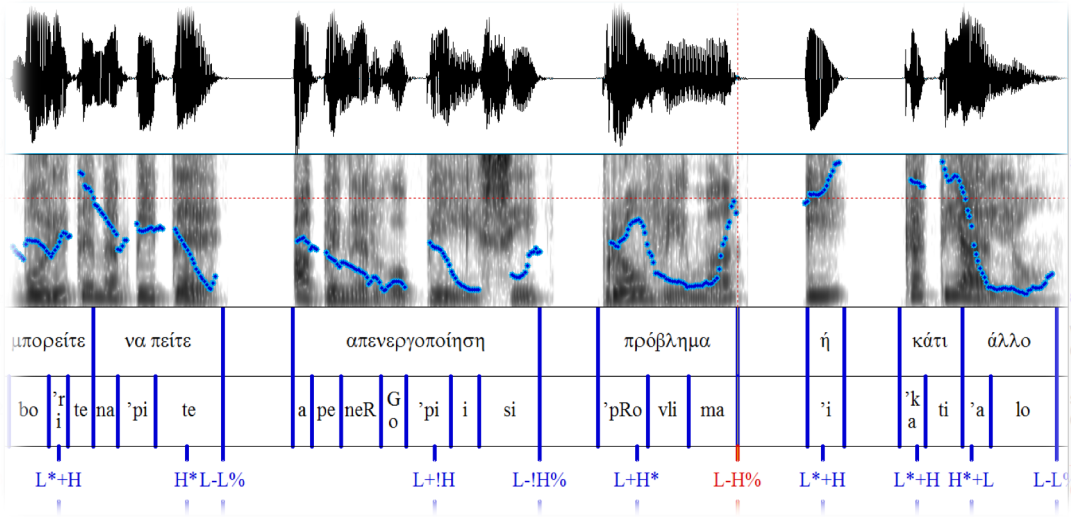

Figure 3 Lists in declaratives; non final list elements are produced with a high edge tone. Note also that the functional word "or" is prosodically marked, produced within its own intonational phrase.

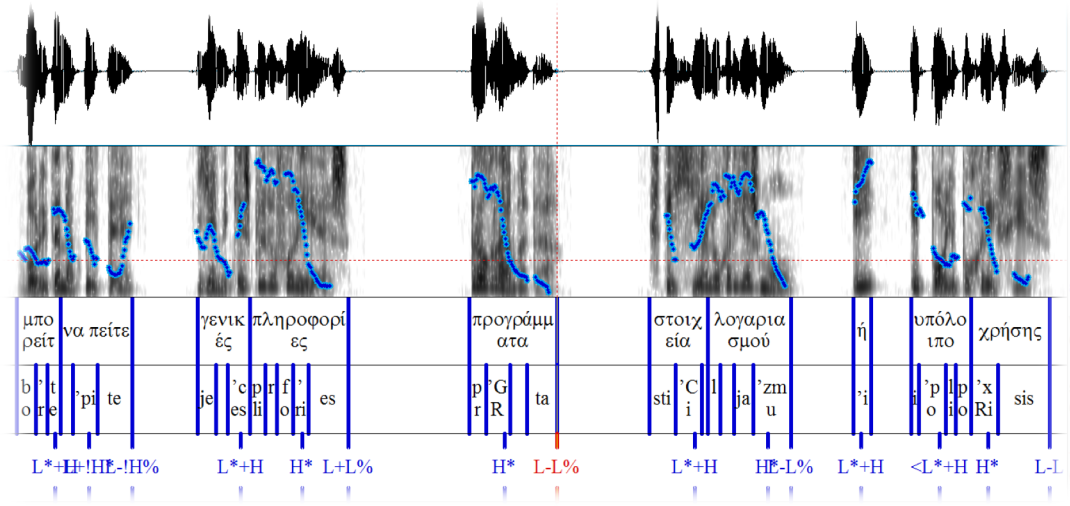

Figure 4 Lists in declaratives; non final list elements are delimited by a low boundary tone.

4.1.2 Negation

Our analysis mostly confirms the results of previous studies (Arvaniti and Baltazani, 2005; Baltazani 2002; 2006) on negation. In 81% of negative declaratives, focus was placed on the negation particle /ðen/, which accordingly carried the nuclear pitch accent, while the subsequent phrase was deaccented. The most commonly used NPA was the L*+H accent followed by the L+H* accent (64,7% and 35,3% respectively). Negative declaratives ended with a L-!H% edge tone or more frequently with a L-L% one (72,4% and 18,6% respectively).

Contrary to what has been shown in previous literature, deaccenting of elements following the negation particle did not always result in the complete elimination of pitch accents in the post nuclear domain. In contrast, we found instances of reduced pitch accents surfacing in the post nuclear domain, corresponding to a bitonal L+!H* accent produced within a compressed pitch range. Figure 5 illustrates a L+!H* post nuclear accent following the focused negation particle ("δεν"). The post nuclear accent is aligned with the word /pera'zmenes/ ("past"). The example corresponds to a case of free second occurrence focus (Büring, 2006; Beaver et al., 2007), where the word /pera'zmenes/ is given and contrasted to other alternatives made available from the discourse context (i.e. past dates are contrasted to future dates).

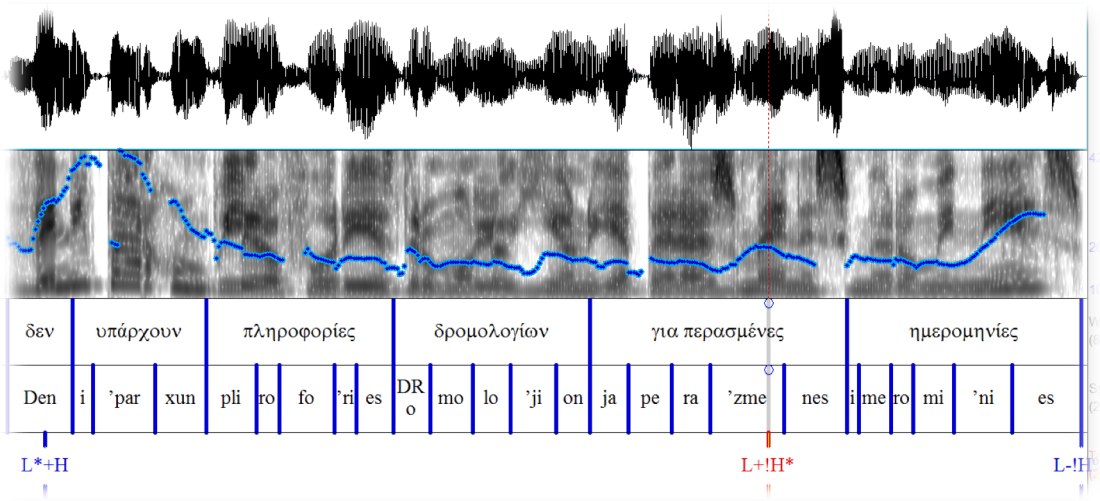

Figure 5 An example of post nuclear accent in negatives. The post nuclear accent aligns with the word "past" (/pera'zmenes/). The speaker states that there is no information available on past dates (as opposed to future dates).

4.2 Wh-Questions

The Wh-Question contour was similar to the one used for negation. The NPA was - in this case - associated with the wh-operator and the subsequent phrase was deaccented. As with negation, the NPAs used were the L*+H and L+H* with the former being the most frequent one (76,7% and 23,3%). Utterances ended in a L-L% or L-!H% boundary (57,6% and 42,4% respectively). Finally, instances of reduced post-nuclear L+!H* accents were reported in the case of Wh-Questions as well (Figure 6).

4.3 Polars

The typical melodies for polar questions were: a) a L* NPA followed by a !H-L% phrase/boundary tone combination (66,7%), and b) a L* NPA followed by a H-L% tone (33,3%)[2]. In our corpus the *non* downstepped H-L% tone combination occurred in cases of early focus, in which the low NPA occurred early in the utterance allowing for more "space" for the canonical realization of the high (H-) phrase tone.

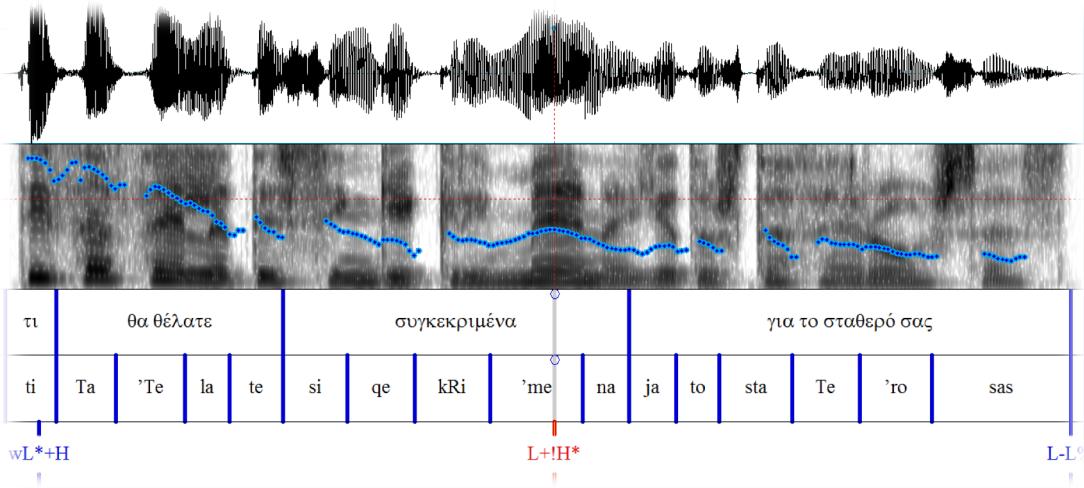

Figure 6 Wh-question contour. The NPA aligns with the wh-word "τι". However, there is a clear post nuclear accent on the word "specifically" (/siqekRi'mena/). The utterance is intended as a clarification question to vague user responses regarding landline phones ("what specifically would you request about...").

[2] This is in accordance with previous studies on polar question intonation (Baltazani and Jun,1999; Arvaniti et al., 2006; Arvaniti, 2009), which describe the polar question tune as consisting of a low pitch accent on the focused word followed by a rise and fall associated with the edge of the utterance. There is still no consensus on the exact nature and representation of this edge movement. Further discussion is beyond the scope of this paper, and the interested reader is referred to Arvaniti (2007) for an in depth analysis.

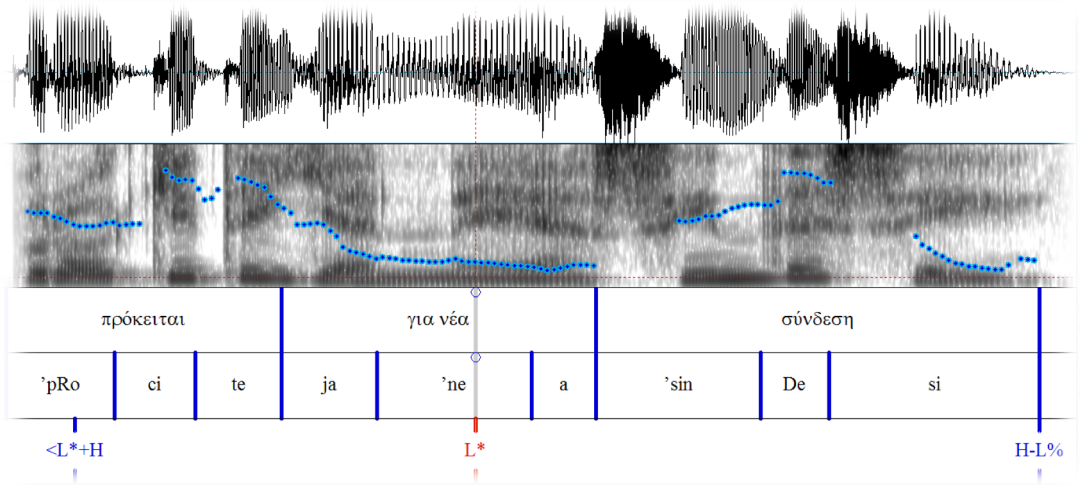

Figure 7 Early focus in polar questions. The L* aligns with the word "new" (/nea/) providing more space for the canonical realization of the H- phrase tone. Note that the phrase tone aligns with the last accented syllable if available - that is in cases of early focus where the subsequent material does not carry sentence stress.

4.3.1 Polars - Lists (Exclusive or)

In lists in in polar questions, the NPA was consistently aligned with the operator "or", which introduces the final list element. The rest of the phrase was deaccented, ending in a L-L% (55,2%) or a L-!H% (44,8%). The most frequent pitch accent associated with the "or" operator was the L*+H accent (67,2%), while a L+H* accent was used in all remaining instances (32,8%). Both final and non final list elements were produced within their own intonational phrase. Non final elements were realized with a high boundary tone (H-H% (28,6%) and L-!H% (1,6%)) or - more frequently - with a polar question melody ending in H-L% (26,79%) or !H-L% (41,07%). Overall, there were two main strategies for the realization of lists in polars:

1. Non final list elements were realized with a polar question melody; the final element ended in a L-L% or L-!H%% and the nuclear pitch accent aligned with the “or” particle (69,8%). An example of this strategy is given in figure 8.
2. Non final list elements were realized with a high boundary; the final element ended in a L-L% or L-!H% and the nuclear pitch accent aligned with the “or” particle (30,2%). An example of this strategy is given in figure 9.
3.

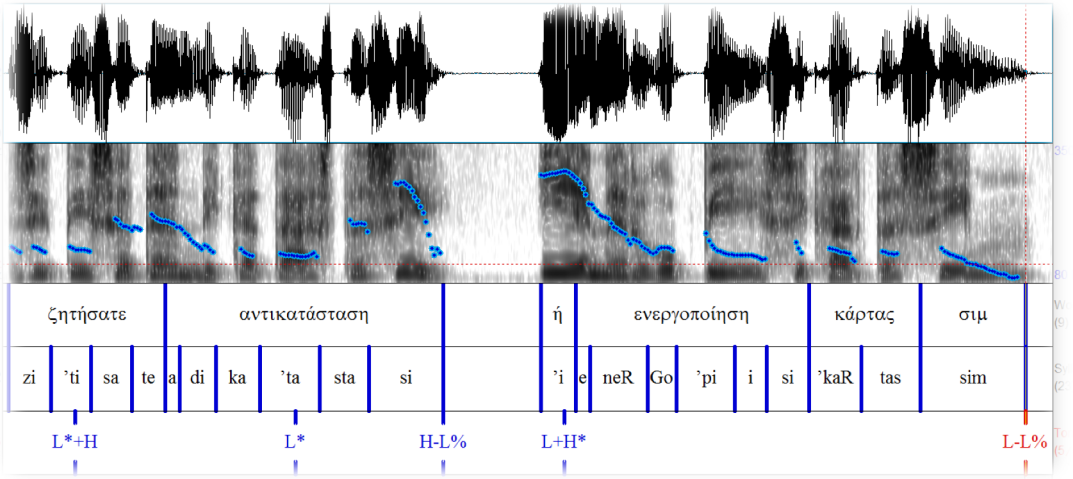

Figure 8 Lists in polars (exclusive or); non final elements produced with a polar question melody.

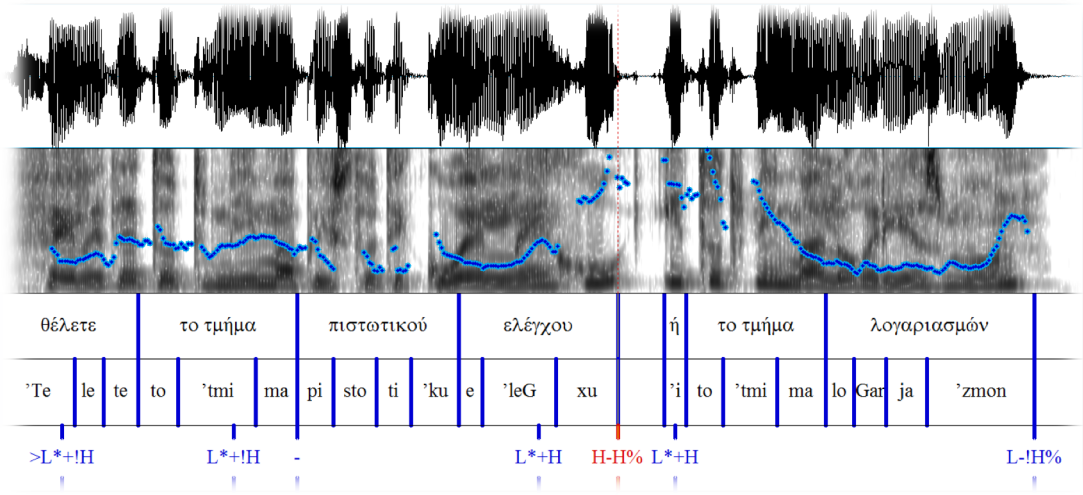

Figure 9 Lists in polars; non final elements are demarcated by a high boundary.

The second strategy was mostly employed by actor C, who, however, consistently resorted to the polar question melody (first strategy) in cases of early focus. In general, the H-H% mid-utterance boundary tone mostly occurred in cases of lists consisting of only two elements. Tables 4 and 5 present the detailed distribution of melodies for non final and final elements respectively. Furthermore, as with wh-questions and negation, instances of pitch accents were observed in the post focal domain, following the "or" operator, and produced with a compressed pitch range (Figure 10).

Table 4 Distribution of phonological melodies for non final list elements in polars

	L* H-L%	L* !H-L	L*+H H-H%	L* H-H%	L+!H* L-H%
Actor A	7	0	1	1	1
Actor B	7	4	3	0	1
Actor C	6	0	5	15	0
Actor D	10	42	3	4	0
TOTAL	30	46	12	20	2
TOTAL %	27.93%	41.44%	10.81%	18.02%	1.80%

Table 5 Distribution of phonological melodies for final list elements in polars

	L+H* L-L%	L*+H L-L%	L+H* L-!H%	L*+H L-!H%
Actor A	4	3	0	0
Actor B	0	0	0	8
Actor C	13	3	0	0
Actor D	4	10	2	20
TOTAL	21	16	2	28
TOTAL %	31.34%	23.88%	2.99%	41.79%

On a final note, it should be made clear that both realization strategies described above apply in the case of the exclusive "or" operator alone. It is therefore important to distinguish between the two different uses (exclusive and inclusive) of the “or” operator, as they: a) elicit different possible answers, b) have distinct prosodic realizations.

More specifically, in the case of exclusive "or" the interlocutor is prompted to choose only one list element. In the case of inclusive "or" the interlocutor may choose one element or both. Examples (1a) and (1b) illustrate the difference. In (1a) - exclusive or - the only acceptable responses are "activation" or "replacement". In (1b) the most likely response is "yes/no".

(1a) S: Are you interested in replacing **or** activating your sim card?
U: Activating it.

(1b) S: Are you interested in replacing or activating your sim card?
U: Yes / Replacing it, yes.

Furthermore inclusive "or" is produced with a typical polar melody question (L* !H-L%) contrary to exclusive "or", in which case the NPA is aligned with the operator and the phrase ends with a L-L% or L-!H% boundary (Cf. figures 8 and 11).

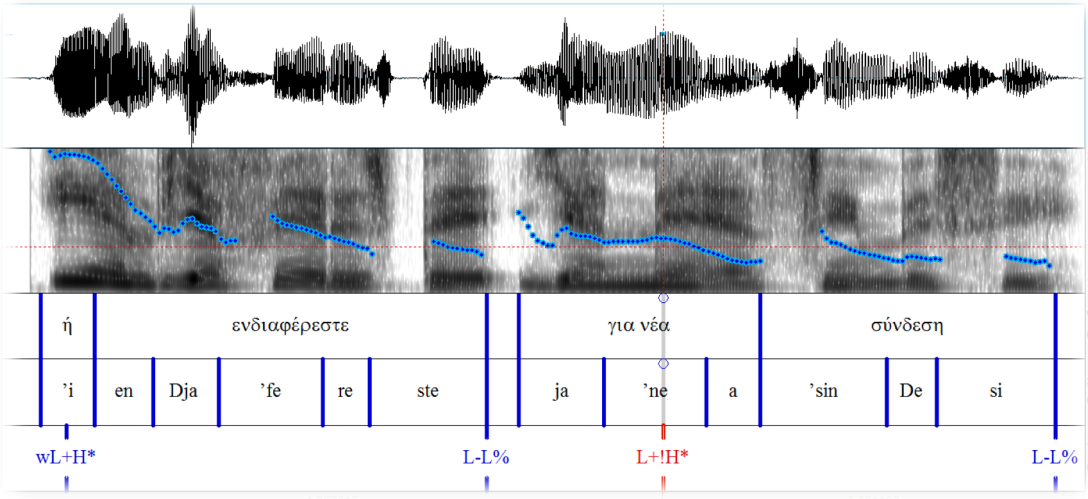

Figure 10 Post nuclear pitch accents in lists in polars (Do you already have an active connection OR are you interested in a NEW one?).

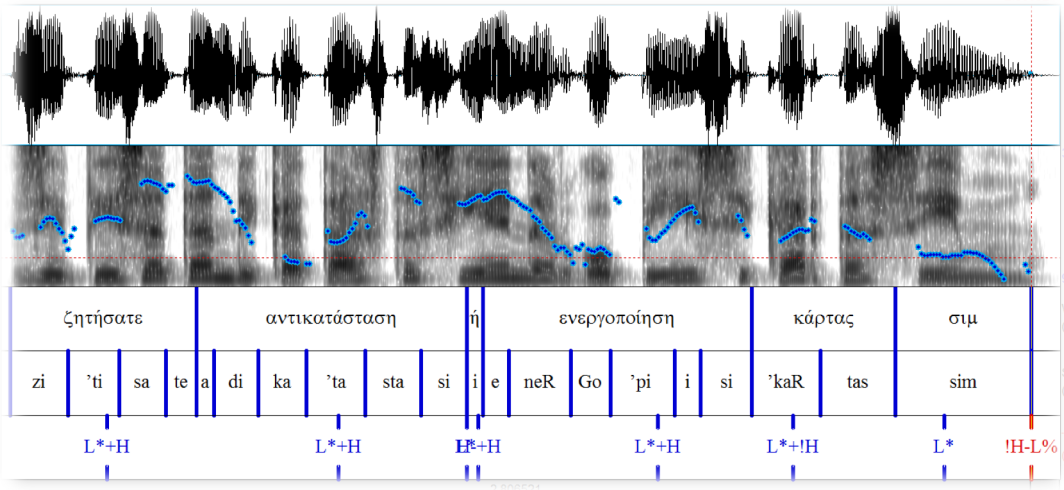

Figure 11 Inclusive "or"

It should be noted that the distinction between exclusive and inclusive "or" made here does not correspond to the prototypical sense of the logical disjunction. In contrast, it is primarily pragmatically determined, in the sense that the two different structures refer to distinct discourse models constructed by the interlocutors. To take an example, in the sentence "Are you interested in activating or deactivating the service?" the arguments "activate" and "deactivate" are logically mutually exclusive (one can do the one or the other, but not both). Therefore, this instance would correspond to the exclusive logical "or". Pragmatically, though, when the above sentence is uttered with a simple polar melody intonation, the two arguments are in fact equated to a single argument, a unary set. In this sense, the syntactic list structure is actually - from a pragmatically motivated point of view - a "pseudolist", as the two elements are considered as one with regards to the dialogue and the requirements of the task at hand.

5. TtS Evaluation Results

In general all three TtS systems were successful at rendering plain declaratives, imperatives and lists in imperatives and declaratives (when provided with appropriate punctuation).

With regards to polar questions, only one system produced three out of four examples with appropriate polar intonation. The rest of the systems produced polars with a high boundary tone similar to the one used for the continuation rise contour.

All systems failed to produce appropriate, grammatically acceptable melodies for negation, wh-questions and lists in polars. More specifically, they all failed to deaccent the material following the respective operator (negation, wh, or operator). Figure 12 illustrates the ill formed contour.

Finally, no system could deduce and handle non default, early focus position.

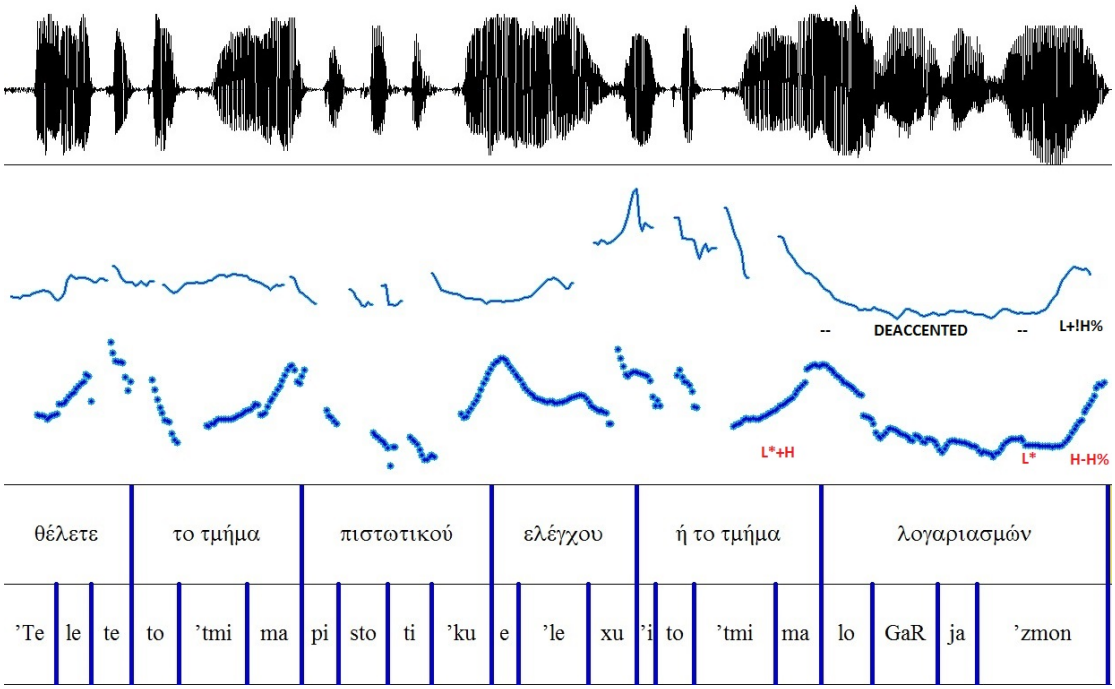

Figure 12 Example of ill formed TtS rendition. Dotted lines represent the F0 contour generated by the TtS system. Solid lines represent a reference contour produced by a voice actor. Note for example the L*+H accent on the word /'tmima/ ("τμήμα") which should have been deaccented.

6. Discussion and Conclusions

Distribution of sentence types in automated spoken dialogue systems revealed a significant percentage of structures (a comparatively high frequency of questions, imperatives, list structuresn and structures with early focus position), which are not common in generic corpora that are normally used for developing databases for building speech synthesizers. TtS evaluation results confirmed this divergence, as all systems failed to generate acceptable contours for wh-questions, lists in polars, and negation, in which case focus is typically associated with non-final position.

On the other hand, the distribution of sentence types was in line with both the nature of task oriented dialogues in general, and human-machine dialogues in particular, where the automated system typically initiates the dialogue, asking questions or providing options for the user to choose from, in order to collaboratively fulfill a task. It is therefore important to take into account the exact nature of the domain and the application when designing the recordings database for developing the Test to Speech system (Black, 2006). For automated dialogue systems, it is important to ensure sufficient coverage for prosodic events related to list structures, questions and non default focus placement among others.

The latter applies to the development of the prosodic module as well, regardless whether simple rule based or data driven techniques are used. The prosodic analysis results may provide insight into the rules and features that could be used for prosodic modeling. In addition to the requirements regarding basic sentence types, at minimum we need to further account for: a) information structure partition and non default focus position, b) list structures and exclusive and inclusive “or”.

Lists in particular were shown to consistently associate with specific realization strategies. It is argued that this consistency, this regularity in their production is due to the strong need for recognizing list structures as means for effective turn taking and conversational sequencing (Selting, 2007). As a rule, the prosodic pattern commonly highlights the distinction between final and non final elements. Non final elements are realized with similar, "parallel" contours, which most of the times contrast with the tune of the final element, which so denotes the closure of the list.

Furthermore the distinction between exclusive and inclusive "or" was shown to greatly influence the tune of the list question as well as the elicited responses, consequently affecting the flow of the dialogue and the effectiveness of the interaction.

It is also important to note that wh-questions, negation, and phrases introduced with the exclusive "or" share the same phonological properties. Taking this into account could reduce the necessary rules for prosody specification, features or number of adequate instances of the relevant prosodic events in the recordings database.

Another interesting outcome of the prosodic analysis was the presence of pitch accents in the post nuclear domain. The possibility of post nuclear pitch accents is not accounted for in the initial autosegmental model of Beckman and Pierrehumbert (1986) where the nuclear pitch accent is defined as the last (leftmost or rightmost) accent in the phonological phrase, and so by definition the existence of post nuclear accents is ruled out. However, in our corpus post nuclear accents were reported in cases of emphasis and relative contrast, such as free second occurrence focus (cf. Section 4.1.2). Our results are corroborated by other recent studies (Norcliffe and Jaeger, 2005;Arvaniti, 2009 on greek polars) which also report instances of compressed post nuclear pitch accents, providing evidence that deaccenting does not involve a complete elimination of prominence in the post nuclear, post focal domain, but rather there are subtle phonetic variations cueing patterns of prominence, which should therefore be included in a grammar model of prosodic structure.

The notions of focus domain, relative emphasis and contrast, as well as sentence types and dialogue acts are incorporated in the preliminary version of a pragmatic annotation schema illustrated in Figure 13. The highest level constituent is a dialogue turn, which is subsequently broken down into utterances and information structure domains (focus-background). Each word is associated with an emphasis level, and the word with the highest emphasis level in the focus domain carries the nuclear pitch accent. This meta-information is intended as input for the speech synthesizer guiding prosody specification.

```
<?xml version="1.0" encoding="UTF-8" ?>
<turn>
<utterance  dialogueAct="InfoRequest,Directive" sentenceType="POLAR">
<domain type="background">
<word emphasisLevel="1"> Πρόκειται </word>
<word emphasisLevel="0"> για </word>
</domain>
<domain type="focus">
<list>
<listItem>
<word emphasisLevel="2"> νέα </word>
<word emphasisLevel="1" punct=","> σύνδεση </word>
</listItem>
<listItem>
<word emphasisLevel="2"> υπάρχουσα </word>
<word emphasisLevel="1" punct=","> σύνδεση </word>
</listItem>
<listItem>
<word emphasisLevel="2" operator="OR_EXCL"> ή </word>
<word emphasisLevel="1"> κάτι </word>
<word emphasisLevel="1" punct=";"> άλλο </word>
</listItem>
</list>
</domain>
</utterance>
</turn>
```

Figure 13 Example pragmatic annotation schema; XML output for the sentence "Is it about a new connection, an existing one, or something else?"

References

Arvaniti, Amalia. 2007. "Greek Phonetics: The State of the Art." *Journal of Greek Linguistics* 8: 97-208.

Arvaniti, Amalia. 2009. "Greek intonation and the phonology of prosody: polar questions revisited." Proceedings of the 8th International Conference on Greek Linguistics, pp. 14-29.

Arvaniti, Amalia, and Mary Baltazani. 2005. "Intonational Analysis and Prosodic Annotation of Greek Spoken Corpora." In Jun, S-A (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press, 84-117.

Arvaniti, Amalia, D. Robert Ladd and Ineke Mennen. 2006. "Tonal Association and Tonal Alignment: Evidence from Greek Polar Questions and Emphatic Statements." *Language and Speech* 49.421-450.

Baltazani, Mary, 2002. "Quantifier scope and the role of intonation in Greek." Unpublished PhD thesis, UCLA.

Baltazani, Mary. 2006. "Intonation and pragmatic interpretation of negation in Greek". *Journal of Pragmatics* 38 (10): 1658–1676.

Baltazani, Mary and S.-A. Jun. 1999. "Focus and topic intonation in Greek." Proceedings of the XIVth International Congress of Phonetic Sciences 2:1305-1308.

Beaver, David, Brady Clark, Edward Flemming, Florian Jaeger, and Maria Wolters. 2007. "When semantics meets phonetics: acoustical studies of second occurrence focus." *Language* 83(2): 245-276.

Alan. W. Black. 2006. "Multilingual speech synthesis." In T. Schultz & K. Kirchhoff (Eds.), *Multilingual speech processing*. (pp. 207-31). Burlington, MA: Elsevier Academic Press.

Boersma, Paul and D. Weenink. 2005. "Praat: doing phonetics by computer (Version 4.3.01)". Retrieved from http://www.praat.org/

Büring, Daniel. 2006. "Been there, marked that—A tentative theory of second occurrence focus." Manuscript, UCLA

Büring, Daniel. 2007. "Semantics, Intonation and Information Structure." In Gillian Ramchad and Charles Reiss (eds), *The Oxford Handbook of Linguistic Interfaces*. Oxford University Press.

Büring, Daniel. 2010. "Towards a typology of focus realization." In:Zimmermann, Malte and Caroline Féry, Eds. *Information Structure. Theoretical, Typological, and Experimental Perspectives*. Oxford University Press. 177-205.

Fellbaum, Klaus, and Georgios Koroupetroglou. 2008. "Principles of Electronic Speech Processing with Applications for People with Disabilities." *Technology and Disability*, 20(2): 55-85.

McKeown, Kathleen R., and Shimei Pan. 2000. "Prosody modelling in concept-to-speech generation: methodological issues." Phil. Trans. R. Soc. Lond. A 358(1769) 1419-1431.

Norcliffe, Elisabeth, and T. Florian Jaeger. 2005. "Accent-free Prosodic Phrases? Accents and Phrasing in the Post-Nuclear Domain." Proceedings of Interspeech 2005.

Selting, Margaret. 2007. "Lists as embedded structures and the prosody of list construction as an interactional resource." *Journal of Pragmatics* 39: 483-526.

Spiliotopoulos, Dimitris, Georgios Petasis, and Georgios Kouroupetroglou. 2008. "A Framework for Language-independent Analysis and Prosodic Feature Annotation of Text Corpora." *Lecture Notes in Artificial Intelligence*, 5246: 517-524.

Stavropoulou, Pepi. 2013. "On the Status of Contrast. Evidence from the Prosodic Domain." *Interdisciplinary Studies on Information Structure* 17 (2013): 1–32 F.Bildhauer and M. Grubic (eds.)

Stavropoulou, Pepi, Dimitris Spiliotopoulos, and Georgios Kouroupetroglou. 2010. "Design and Development of an Automated Voice Agent: Theory and Practice Brought Together." Chapter in the book: *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*, 2010, Information Science Reference Press (IGI Global), Pennsylvania, USA.

Syrdal, Ann, and Kim Yeon-Jun. 2008. "Dialog speech acts and prosody: Considerations for TTS." In Proc. of the Speech prosody, Brazil.

Taylor, Paul. 2000. "Concept-to-speech synthesis by phonological structure matching." Phil. Trans. R. Soc. Lond. A 358(1769): 1403-1417.

Xydas, Gerasimos, Dimitris Spiliotopoulos, and Georgios Kouroupetroglou. 2003a. "Prosody Prediction from Linguistically Enriched Documents Based on a Machine Learning Algorithm." In proceedings of the 6th International Conference of Greek Linguistics (6th ICGL), Rethymno, Greece, September 18-21 2003.

Xydas, Gerasimos, Dimitris Spiliotopoulos, and Georgios Kouroupetroglou. 2003b. "Building Prosodic Structures in a Concept-to-Speech System." In proceedings of Workshop on Balkan Language Resources and Tools, 1st Balkan Conference on Informatics (BCI-2003), Thessaloniki, Greece, November 21, 2003. http://speech.di.uoa.gr/sppages/spppdf/web-xydas_balkan2003.pdf

Xydas, Gerasimos, Dimitris Spiliotopoulos, and Georgios Kouroupetroglou. 2004. "Modeling Prosodic Structures in Linguistically Enriched Environments." Lecture Notes in Artificial Intelligence, 3206: 521-528.

Xydas, Gerasimos, Dimitris Spiliotopoulos, and Georgios Kouroupetroglou. 2005. "Modelling Improved Prosody Generation from High-Level Linguistically Annotated Corpora." IEICE Trans. Inf. & Syst., Special Issue on Corpus-Based Speech Technologies, E88-D(3): 510-518.

Young, S.J. and F. Fallside, F. (1979) “Speech synthesis from concept: A method for speech output from information systems.” J. Acoust. Soc. Am. 66: 685-695.