In Proceedings of the 3rd Hellenic Conference on Artificial Intelligence (SETN04), Samos, Greece, May 5-8, 2004

Text Normalization for the Pronunciation of Non-Standard Words in an Inflected Language

Gerasimos Xydas, Georgios Karberis and Georgios Kouroupertroglou

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
Speech Group
{gxydas, grad0350, koupe}@di.uoa.gr

Abstract. In this paper we present a novel approach, called "Text to Pronunciation (TtP)", for the proper normalization of Non-Standard Words (NSWs) in unrestricted texts. The methodology deals with inflection issues for the consistency of the NSWs with the syntactic structure of the utterances they belong to. Moreover, for the achievement of an augmented auditory representation of NSWs in Text-to-Speech (TtS) systems, we introduce the coupling of the standard normalizer with: i) a language generator that compiles pronunciation formats and ii) VoiceXML attributes for the guidance of the underlying TtS to imitate the human speaking style in the case of numbers. For the evaluation of the above model in the Greek language we have used a 158K word corpus with 4499 numerical expressions. We achieved an internal error rate of 7,67% however, only 1,02% were perceivable errors due to the nature of the language.

1 Introduction

Unrestricted texts include Standard Words (Common Words and Proper Names) and Non-Standard Words (NSWs). Standard Words have a specific pronunciation that can be phonetically described either in a lexicon, using a disambiguation processing to some extent, or by letter-to-sound rules. By definition, NSWs comprise numerical patterns and alphabetical strings that do not have a regular entry in a lexicon and their pronunciation needs to be generated by a more complicated natural language process. In inflected languages word sequences that result from NSWs need to be proper inflected and converted into the right gender in order to match the syntactic structure of the sentences and the target noun they refer to. Even so, there are still some Text-to-Speech (TtS) oriented issues concerning the style, the rate and the format of the pronunciation of NSWs that have not been addressed yet. For example, humans tend to read out long numbers slowly and with pauses between groups of digits.

Most of the previous works deal with NSWs' pronunciation in Text-to-Speech systems, however, NSWs constitutes a problem in the fields of information retrieval and speech recognition [6]. Most of the proposed approaches are language specific as the problem depends on language properties. Even so, there are some issues, like the inflection of the NSWs, which have been partially solved. For example, in the

German language there are two systems that deal with normalization: Bell Labs TtS [1] and FELIX [2]. FELIX analyzes the text syntactically using the Zingle [3] algorithm and the utterance pronunciation is determined by the Part-of-Speech information. In the Bell Labs approach, there is an attempt here to deal with ambiguities but the lack of syntactical analysis limits the capabilities of the system. In the Japanese language [4] every possible role of a word in a sentence is scored and, after analysis, the role with the highest dependency score is selected. Thus, the appropriate pronunciation is applied for the normalization of the NSWs. A model that uses a pre-processor performing syntax analysis of sentences was presented in [5] for English. Though it is presented to be a language independent solution, there is no care for inflections.

The Johns Hopkins University Summer Workshop (WS99) research project [6] made a systematic effort to build a general solution of the NSW's normalization problem in the English language. Later on, this was applied to the Asian languages in [7]. The application of this model to the Greek language has the major drawback of the inflection.

In the "Text to Pronunciation" (TtP) work we deal with three important elements in the normalization of texts. The first is the dynamic definition of the pronounceable format of NSWs through a Language Generator model, leading to increased semantics in the synthesized speech. The second deals with the inflection of word lattices that are generated during the expansion of NSWs, so that normalized expressions are consistent with the syntax structure of the utterances they belong to and to ensure the sequence of tenses and genders in nominal phrases. Other important issues that have not been addressed before and we accommodate in this work are the capability of inserting SSML [10] (or VoiceXML or any other speech markup language, SAPI etc) tags mainly for defining short breaks between groups of digits in cases of long numbers.The rest of the paper focuses on the numerical's problem, which is (a) more general, (b) more important and (c) shares similar, if not more complex, methods to the alphabetical cases. Thus, the same methodology can be and has been applied to alphabetical expressions as well.

2. The TtP model

Figure 1 presents the block diagram of the TtP model. The individual components and the data used by them can by either domain specific (e.g. in economical texts there are several numerical patterns that have a different pronunciation in the sports domain) or generic. The first yields better performance and supports disambiguation.

2.1 Tokenizer, Splitter and Classifier

These have been described in [6]. Their functionality differs in Greek but the model works similar. The main purpose of the Tokenizer is to successfully identify End of Sentences (EoS) and to create tokens from the given sentence. We use the EoS of the Greek TtS system DEMOSTHeNES [8]. In cases of dot punctuation, two

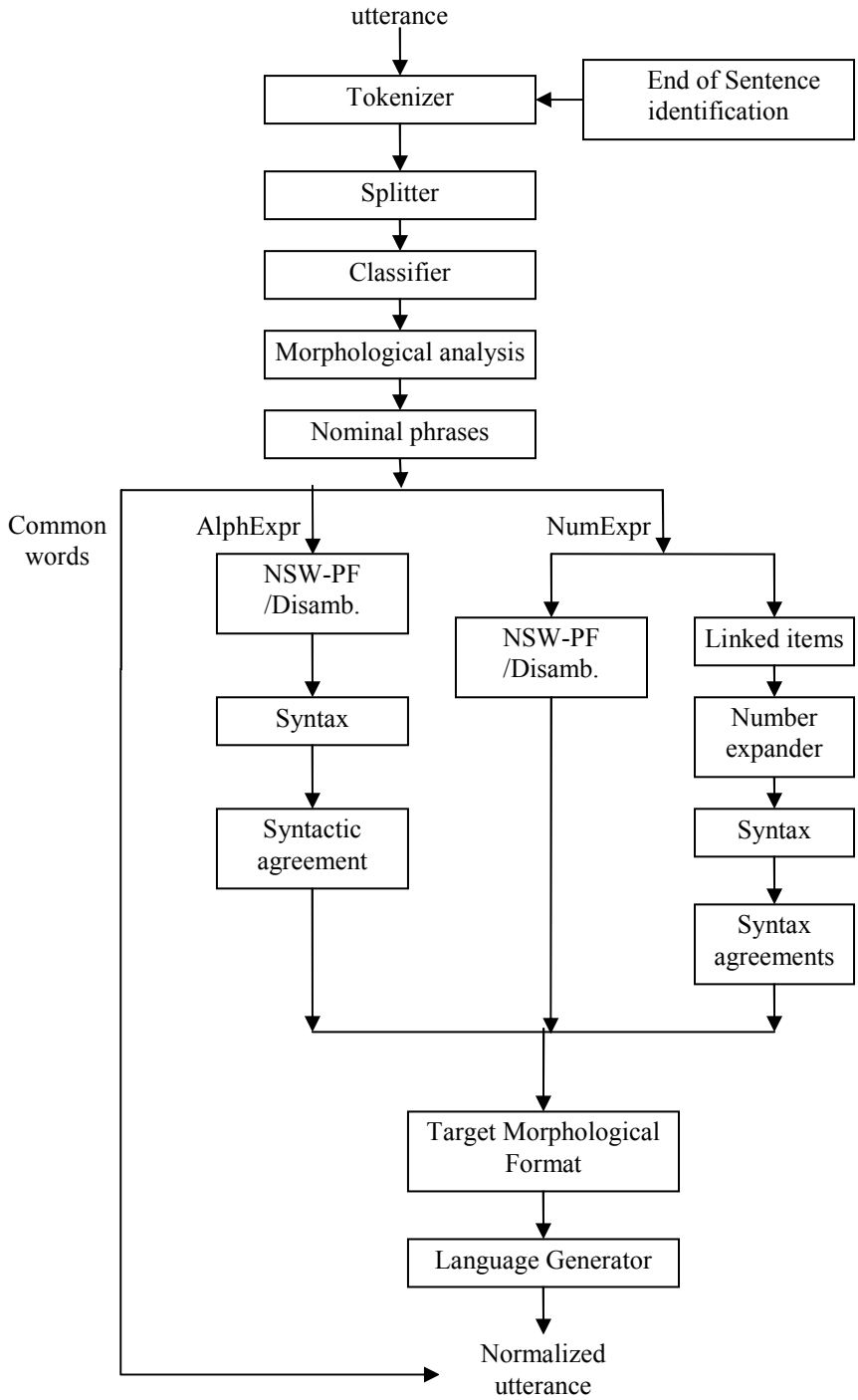

Fig. 1. The architecture of the NfP.

lists are parsed for `EoS` disambiguation: acronyms (`"\\([Α-Ω]\\.\\)*[Α-Ω]\\.?"`) and a list of abbreviations. We consider them as `EoS` if the next token starts with a capital letter and it is not an Out-of-Vocabulary (OOV) word (i.e. likely to be a proper name). For example "ο κ. Νικολάου" and "το Ι.Κ.Α. Αθήνας" are not `EoS`. This is not optimum; however, the evaluation showed that it does not affect the model for the specific task (2 errors in 4499 cases).

The role of splitter is to split tokens that are not pure numerical or alphabetical expressions. In inflected languages there are some more issues to address. For example "25χρονος" can not be split into "25" and "χρονος" and "3-4 ώρες" can be split into "3", dash, "4" and "ώρες", but with a mark, because even if they constitute different tokens, both numbers should be inflected and converted into the gender that matches the common noun "ώρες". These are being handled in the Language Generator component.

Finally, the classifier is based on Regular Expressions and identifies (a) in case of numerics: cardinal, ordinal, dates, hours, time, years, telephone, currencies, percentages, postal codes, street number, IP addresses, digits and (b) abbreviations: to expand, as words and as letters.

2.4 Expansion of NSWs

Firstly, we perform the expansion on the alphabetical expressions and then on the numerical ones. Thus, abbreviations that characterize numerical expressions (e.g. "5 δις.") are normalized prior to these numerics, so that the numerics would be able to inherit the morphological structure on the expanded abbreviation.

2.5 The Non-Standard Word Pronounceable Format

One of the main aspects of this work is the definition of the NSW Pronounceable Format (NSW-PF) that might accommodate any kind of NSW. This format allows a flexible control over the way numbers and strings are pronounced. It assigns to an expression a digit or alpha grouping and a corresponding Target Morphological Format (TMF). Table 1 shows some examples. Symbols on the "Regular Expressions" field should have a matching symbol in the "NSW-PF". A sequence of similar symbols in NSW-PF defines a number of that many digits as the number of symbols in the sequence (e.g. ## defines a 2-digit number) and this corresponds to the same number of digits in the Regular Expression. Thus, wildcards and ranges of digits (e.g. "[0-9]*") are not allowed in this specification.

Reg. Expr.	Class	NSW-PF
##/#/####	Date	<**tmf** gender="neutral" case="nominative" number="singular">**##**</**tmf**> <**tmf** gender="neutral" case="genitive" number="singular">**#**</**tmf**> <**tmf** gender="feminine" case="nominative" number="singular">**####**</**tmf**>
###-#######	Tel.	<**ssml**:prosody rate="-20%"> <**tmf** gender="neutral" case="nominative"

		number=”singular”>**#**</**tmf**> <**tmf** gender=”neutral” case=”nominative” number=”singular”>##</**tmf**> <**ssml**:break time=”long”/> <**tmf** gender=”neutral” case=”nominative” number=”singular”>##</**tmf**> <**ssml**:break/> <tmf gender=”neutral” case=”nominative” number=”singular”>##</**tmf**> <**ssml**:break/> <**tmf** gender=”neutral” case=”nominative” number=”singular”>#</**tmf**> <**ssml**:break/> <**tmf** gender=”neutral” case=”nominative” number=”singular”>##</**tmf**> </**ssml**:prosody>
##ωρος	AlfNum Stem	<**tmf** gender=”masculine” case=”nominative” number=”singular” type=”alphanum_stem”>##</**tmf**>ωρος
στις #/#	Date	στις <**tmf** gender=”neutral” case=”nominative” number=”singular”>#</**tmf**> <**tmf** gender=”neutral” case=”genitive” number=”singular” type=”ordinal”>#</**tmf**>
το #/#	Ord.	το <**tmf** gender=”neutral” case=”nominative” number=”singular”>#</**tmf**> <**tmf** gender=”neutral” case=”nominative” number=”singular” type=”ordinal”>#</**tmf**>

Table 1. Examples of the Non-Standard Pronounceable Format. In the *telephone* example SSML tags have been incorporated to be parsed latter by the TtS system.

where # = 0..9.

The NSW Pronounceable Format allows five things to happen:

1. To deal with ambiguities, like the last two numerics on the above table.
2. To mix plain text with annotated text in the format section.
3. To segment a number to smaller groups of digits so that it could be pronounced in a relaxed and understandable way (imagine reading out the number “456675342345” as a single entity).
4. To define the morphology of the sub-groups in specific cases and domains, where the user ensures the right pronunciation of a NSW in terms of the TMF. According to the above definitions, the telephone number “210-7275320” will be pronounced as “two ten seventy two seventy five three twenty”. The NSW-PF grammar has some restrictions itself, but these are not evaluated here and are out of the scope of this paper.
5. To incorporate SSML (or VoiceXML or any other speech markup, SAPI etc) tags in the NSW-PF so that, for example, the TtS is forced to pause in specific points, like humans do when reading long numbers. The above telephone example will be read out in a 20% decreased rate (which is a natural behavior; humans speaks slow in order for the number to be easier to be memorized by the interlocutor) and with medium and long pauses between the group of digits: “two ten// seventy two/ seventy five/ three/ twenty”.

2.6 The Target Morphological Format

The Target Morphological Format (TMF) has been introduced to allow the definition of the pronunciation format of NSWs. The TMF entries have an XML form (the <tmf> entity in Table 1) and might include as attributes any field from the morphological lexicon. In the Language Generator component the enclosed words will be converted in order to conform to these directives.

2.7 The Number Expander

In cases of numerical expressions that are not defined in the NSW-PF, we use a dynamic Number Expander that differs from the Language Generator as the linguistic nature of numbers can be characterized as "non-normal". Numerics have three states: cardinal, ordinal and alphanumeric (e.g. "14ωρο"). Consider the string "των 1636 ανθρώπων". After all the above steps and dealing with the exceptions we will generate "των {χίλια} {εξακόσια τριάντα έξι} ανθρώπων". The cardinal, ordinal or alphanumeric information is passed to the Syntactic Agreement component for further processing.

2.8 Syntactical analysis

Syntactical analysis is being performed using templates for the Greek language. Each template defines a sequence of POS occurrences and can determine the subject and the object or the link verb complement. Table 2 shows some templates that determines the object. Many of the templates that stand for the object are common for the predicate as well. Though the covering achieved by this approach is small for the richness of the Greek language, however, this seems adequate for the specific problem we deal with: the normalization errors caused by faulty syntax were 4,47% while the noticeable errors were only 0,89%. This is because the syntax information we mainly look for is the identification of nominal phrases that can be predicted by such an approach.

Object
At+No+Pn+Cj+Pn
At+No+Cj+Pn
Pn+Cj+Ar+No+Pn
Pn+Cj+Ar+No
...

Table 2. Syntactic templates (sequences of POS) for the Object.

Syntactical analysis is also important in inflected languages for another reason: assume the text "3-4 ώρες". The splitter will split these in "3", dash, "4" and "ώρες". Both numbers should be inflected and match the gender of "ώρες". There are other cases as well of conjunctive numbers with a common noun. This are dealt by the

following directive {NSW+[lnk+NSW]*+No}, where lnk is a defined set of links between consecutive NSW: "-", ",", "&", "και", "εώς", "ως", "μέχρι", "στα", "στις", "προς", "με", "ή".

2.9 Syntactic Agreement

We collected a number of main rules found in the grammar of Modern Greek [9]. These proved to achieve the desired consistency between numerics and nouns. Cases of syntactic agreement constitute 1635 out of 4499 numerics in the corpus (36,34%) and raised errors otherwise. These agreements are:

2.9.1 Agreement within nominal phrases

The gender, case and number of the noun must agree with the gender, case and number of its specifiers (in our case numbers). Thus, we can determine the gender, number and case of numbers if we know the morphological information of the noun they specify or the determiner of the noun (in case the head-noun itself is missing). For instance, the above example will produce a TMF of:
"των <tmf gender='masculine' number='plural' case='genitive'> χίλια </tmf> <tmf gender='masculine' number='plural' case='genitive'> εξακόσια τριάντα έξι</tmf> ανθρώπων".

2.9.2 Agreement between subject and verb

The verb of a sentence inherits the number and the person of the subject. Moreover, the tokens that constitute the subject are transformed to the nominative case. This rule we deal with cases of ambiguities in the morphology of words that constitute the subject. For example: "Τα 1500 έφτασαν εχθές.". In order to form the number "1500" we look in the nominal phrase "τα 1500". However, "τα" can be either in the nominative or the accusative case. Since this nominal phrase is the subject of the sentence, the case of the tokens in it is chosen to be the nominative.

2.9.3 Object in the accusative case

The object of a sentence is always in the accusative case. Thus, all the tokens that constitute the object are considered to be in the accusative case: "Το μουσείο δέχεται καθημερινά 1500 επισκέπτες".

2.9.4 Predicate in the nominative case

On the other hand, the predicate of a sentence is always in the nominative case. Furthermore, it inherits the case and the gender of the subject, as complement describe or identifies the subject: "Οι επιτυχόντες είναι 1501".

2.10 The Language Generator

The Language Generator component is able to parse sequences of TMFs and generate the word lattices with the corresponding morphological features. For Greek, this is achieved by using of a morphological lexicon that includes (per word):

- Word
- Morpheme information
- Inflection code
- Stem
- Lemma

The Inflection code field corresponds to a template of affixes. Thus, given a word we retrieve its stem and following the template instructions we form the target word. All the inflectional affixes in any applicable gender can be dynamically generated for any word in this lexicon. Table 3 shows some entries of this lexicon, while Table 4 illustrates the corresponding templates.

Word	**Morph**	**Infl. Code**	**Stem**	**Lemma**
εξάρτημα	NoNeSgNm	O49	εξάρτημ	Εξάρτημα
σελίδα	NoFeSgNm	O26	σελίδ	Σελίδα

Table 3. Entries in the morphological lexicon.

Infl. Code	**S1**	**S2**	**S3**	**S4**	**P1**	**P2**	**P3**	**P4**
O49	α	ατος(+1)	α	α	ατα(+1)	ατών	ατα(+1)	ατα(+1)
O26	α	Ας	α	α	ες	ων	ες	Ες

Table 4. Desinence's templates. S1 to S4 and P1 to P4 stand for the singular and plural cases.

The lexicon defines 53 affix templates, covering all Greek nouns and 17 templates for the adjectives and the participles. The total amount of word covering is currently 1.077.458 words. Articles have been encoded separately. Numbers are formed depending whether they are cardinal, ordinal or alphanumerical. Default type is cardinal. Default gender is neutral. Default case is nominative. These "default" assumptions have very impressive effects on the overall evaluation as most of the numerical NSW fail into them.

3. Evaluation

During the evaluation, we distinguish between "system errors" and "perceivable errors". Due to the fact that in cases of weakness to predict or generate the correct normalized form of a NSW the system assumes that the token should be rendered in the corresponding neutral, singular and nominative form, there are errors that are not noticeable, because this assumption stands for the majority of the cases: e.g. only numbers ending in "1", "3" or "4" have different forms in male, female and neutral genders. Thus, system errors refer to "wrong or uninformed prediction but possibly

right pronunciation" while perceivable errors refer to "wrong pronunciation". The size of the corpus used for evaluation (158557 words - 4499 NSWs) verifies this.

Evaluation involves only the standard normalization procedures and not the pronunciation oriented ones. Firstly, we built and hand annotated a corpus in the Speech Group of the University of Athens. The corpus was selected from 4 major on-line newspapers and the subjects covered were:

Category	Total Words	Nums	%
Sports	35994	1225	3,40
Social issues	27825	531	1,91
Politics	29190	482	1,65
Economy	55772	2013	3,61
Science	4826	118	2,45
Weather	3432	130	3,79
TOTAL	158557	4499	2,84

Table 5. Statistics of the NSW in the collected corpus.

Category	Nums	AlphNum		Regular Expr.		Nominal phrases		Discourse	
Sports	1225	357	29,14%	247	20,16%	343	28,0%	278	22,69%
Social issues	531	42	7,91%	203	38,23%	277	52,17%	9	1,70%
Politics	482	71	14,73%	257	53,32%	136	28,22%	18	3,73%
Economy	2013	57	2,83%	1158	57,53%	761	37,80%	37	1,84%
Science	118	5	4,24%	42	35,59%	69	58,47%	2	1,69%
Weather	130	60	46,15%	20	15,38%	49	37,69%	1	0,77%
TOTAL	**4499**	**592**	**13,16%**	**1927**	**42,83%**	**1635**	**36,34%**	**345**	**7,67%**

Table 6. Statistics of the NSWs. The classification of the NSWs is presented here. The last case (Discourse, mainly surface anaphora) could not be currently handled by the system.

The current implementation of the TtP is able to handle in generic or domain specific environments most of the NSWs. Table 6 illustrates the kind of numerical NSWs and how they were classified. The last column shows the actual errors of the model and need discourse analysis to be handled. However, only a 1,02% was actual perceivable errors.

Comparing against a legacy model of DEMOSTHeNES based on the UoA Transcriber (FSA engine), the improvements were dramatics: the UoA Transcriber is able to handle all of the alphanumerics cases, most of the Regular Expressions but there is not any provision for nominal phrases and of course discourse analysis. Thus, the optimum error rate for the UoA Transcriber is 44,0%.

The introduction of the NSW-PF provides improved effects: "2107275320" is not pronounced like "δύο δισσεατομμύρια εκατόν επτά εκατομμύρια διακόσιες

εβδομήντα πέντε χιλιάδες τριακόσια είκοσι" but slowly "δύο δέκα {pause} εβδομήντα δύο {pause} εβδομήντα πέντε {pause} τρία {pause} είκοσι".

4. Conclusions

We presented a novel model for the normalization of NSWs that achieves improved semantics of the synthesized speech by dealing with inflection issues and enhanced auditory representation in cases of NSWs by (a) defining the NSW Pronounceable Format and (b) incorporating VoiceXML attributes to the normalized tokens. The evaluation of the model for the Greek language showed the drastically improvement of 36,33% in correct normalization over a legacy model of the DEMOSTHeNES Speech Composer, while the auditory enhancements have not been evaluated.

Acknowledgements

We would like to thank Pepi Stavropoulou for her significant contribution during the evaluation.

References

1. B. Mobius, R. Sproat, J. van Santen, J. Olive: The Bell Labs German Text-To-Speech system: An overview. In Proceedings of EUROSPEECH '97, Volume IV, p. 2443-2446, (1997)
2. G. Fries and A. Wirth: FELIX – A TTS System with Improved pre-processing and source signal generation. In Proceedings of EUROSPEECH '97, Vol. II, p. 589-592, (1997)
3. H. Zingle: Traitement de la prosodie allemande dans un systeme de synthese de la parole. These pour le 'Doctorat d'Etat, Universite de Strasbourg II, (1982)
4. Y. Ooyama, M. Miyazaki, S. Ikehara: Natural Language Processing in a Japanese Text-To-Speech System. In Proceedings of the Annual Computer Science Conference, p. 40-47, ACM, (1987)
5. D. Coughlin: Leveraging Syntactic Information for Text Normalization. Lecture Notes in Artificial Intelligence (LNAI), Vol. 1692, p.95-100, (1999)
6. Richard Sproat, Alan Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards: Normalization of non-standard words. Computer Speech and Language, 15(3), p. 287-333, (2001)
7. Olinsky, G. and Black, A.: Non-Standard Word and Homograph Resolution for Asian Language Text Analysis. In Proceedings of ICSLP2000, Beijing, China, (2000)
8 Xydas G. and Kouroupetroglou G.: The DEMOSTHeNES Speech Composer. In Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland, August 29th - September 1st, pp 167-172, (2001)
9 Babiniotis, G. and Christou, K.: The Grammar of Modern Greek, II. The verb. Ellinika Grammata, (1998)
10 Burnett, D., Walker, M. and Hunt, A.: Speech Synthesis Markup Language Version 1.0. W3C Working Draft, http://www.w3.org/TR/speech-synthesis