

“My Small Slim Greek ASR System” or Automatic Speech Recognition of Modern Greek Broadcast News

Jürgen Riedler[†], Sergios Katsikas^{*}

[†]SAIL Labs Vienna, Austria

^{*}FASK Germersheim, Johannes Gutenberg-Universität Mainz, Germany

juergen@sail-technology.com

Abstract

In this paper we report on the development of a Modern Greek large-vocabulary continuous-speech recognition system. We discuss lexical modelling with respect to pronunciation generation and examine its effects on word accuracies. Peculiarities of Modern Greek as a highly inflectional language and their challenges for speech recognition are addressed.

1. Introduction

Modern Greek *Koine* or Standard Modern Greek, the official language of Greece and Cyprus, is the latest variety of Europe’s oldest literary language (3500 years), following Mycenaean, Ancient, Hellenistic, and Byzantine Greek. Research objectives within (the ongoing) CIMWOS¹ project [1] comprise *inter alia* a Modern Greek automatic speech recognition (ASR) system. After providing a brief linguistic overview of Modern Greek *Koine* we specify the prerequisites for ASR, which would be: audio recordings with corresponding transcriptions to train acoustic models, text corpora for language modelling, and recognition lexica inclusive pronunciation generation. Finally we disclose word error rates of experiments employing various recognition dictionaries and discuss major problems of lexical and language modelling for a highly inflectional language.

2. Notes on Modern Greek structure [2]

2.1. Phonological system

The phonological system of Modern Greek [3] consists of five vowel phonemes: /a/, /ɛ/, /i/, /o/, /u/ and 20 consonant phonemes: the plosives /p/, /b/, /t/, /d/, /k/, /g/, the fricatives /f/, /v/, /θ/, /ð/, /s/, /z/, /x/, /ɣ/, the affricates /t͡s/, /d͡z/, the nasals /m/, /n/, the lateral /l/ and the apical trill /r/. The most important allophone-generating phonological processes are:

- palatalisation of /k/, /g/, /x/, /ɣ/ to [ç], [j], [ç], [j] before /i/ or /ɛ/
- /k/, /g/, /x/, /ɣ/, /n/, /l/ merge with following glide [j] (non-syllabic allophone of /i/) to palatals: [ç], [j], [ç], [j], [ɲ], [ʎ], e.g. εννιά /eniˈa/ → *[eˈnja] → [eˈɲa]
- sonorisation of /p/, /t/, /k/, /t͡s/ to [b], [d], [g], [d͡z] after /n/, often with denasalisation in informal speech, e.g. τον πατέρα /ton paˈtera/ → [tombaˈtera] or [tobaˈtera]
- regressive assimilation of place of articulation of /n/ to the following consonant

¹Combined **IM**age and **WO**rd Spotting – funded by the Information Society Technologies Programme under contract: IST-1999-12203

- /n/ → [m] before /p/, /b/, see former example
- /n/ → [ŋ] before /k/, /g/, /x/, /ɣ/,

e.g. τον Κόστα /ton ˈkosta/ → [tonˈgosta] or [toˈgosta]

- sonorisation of /s/ to [z] before voiced consonants, e.g. της λέω /tis ˈleɔ/ → [tizˈleɔ]

Within syntactic phrases (e.g. article - noun - possessive pronoun) certain phonological processes usually extend even across word boundaries (see examples above), but only if there is no pause between the words. This can cause homophony of phrases, e.g. [timˈbira] or [tiˈbira] could mean both την μπύρα “the beer {acc.}” or την πήρα “I picked her up/I called her etc.”, and represents an almost inevitable source of word errors in ASR (cf. Section 4.1).

2.2. Prosody

The functional load of prosodic features in Modern Greek is extremely high, since word stress and intonation are highly distinctive. There are hundreds of prime-stress minimal pairs (e.g. πότε “when” vs. ποτέ “never”), stress fulfills various morphological functions [4] and moreover, intonation patterns provide in most cases the only distinction between declarative clauses and yes-no questions (e.g. [o ˈjanis ˈin(ɛ) ɛˈðo\] “John is here” vs. [o ˈjanis ˈin(ɛ) ɛˈðoˀ] “Is John here?”). This is the reason why we introduced word stress as a part of suprasegmental structure into our phone sets, see Section 3.1 and 3.2.

2.3. Morphology

Modern Greek is a prototypical inflectional language, *i.e.* a potentially huge number of different word forms may be derived from one basic stem (lemma). In particular verb inflection is very rich: by combining two stems, three sets of endings, a few modal particles, an auxiliary verb and the participle, every active verb can produce about 200 forms, if we take all syntactically defined categories (three aspects, six moods, eight tenses, *etc.*) into account, despite of (partial) homonymies. This number is twice as big for verbs that exhibit a medio-passive voice, which is formed synthetically. Different verb forms can differ from each other in the ending, in accentuation as well as in the stem (there are also irregular verbs with suppletive roots, e.g. βλέπω [ˈvlepo] “I see” vs. είδα [ˈiða] “I saw”), and finally, active verbs consisting of two syllables have in past tense a sort of prefix (augment) carrying the stress on the antepenultimate syllable. Nouns show, depending on their inflectional class, between 4 and 7 different forms, adjectives about 40 (including comparative and elative). Due to ambiguities of various morphological rules and the bistructurality of Modern Greek (parallel use of old and new forms, depending on the situation [5]), inflectional forms are often unpredictable without the informa-

tion of a dictionary or knowledge of Ancient Greek.

Inflection and the obligatory concord between constituents of sentences draw distinctions unknown in a rather isolating language like English. For example the English sentence

I have an old friend who knows a famous Italian singer.

may have four different meanings (depending on the gender of my friend and the singer), of which we give two for brevity:

Έχω έναν παλιό φίλο που ξέρει έναν γνωστό
Ιταλό τραγουδιστή. (male, male)
Έχω μια παλιά φίλη που ξέρει μια γνωστή
Ιταλίδα τραγουδίστρια. (female, female)

Modern Greek word formation processes [6] are very complex though not very productive. Various mutations of morphemes and bistructurality prevent the predictability of derivatives and compounds. For example, the stems within the verb forms κλέβω “I steel”, έκλεψα “I stole” do not evidently imply those in derivatives like κλέφτης “thief”, κλοπή “theft” or in a compound like κλεπτομανής “cleptomaniac”.

Since syntactic relations between constituents of a sentence are mostly expressed by inflection, Modern Greek constituent order is fairly free (with a few exceptions, e.g. the order of particles and clitics in relation to the verb). Word order has rather a pragmatic than syntactic function (e.g. topicalisation).

It is obvious from the above that inflection as well as syntactical freedom present outstanding demands on lexical and language modelling.

3. Phonetic transcription

Modern Greek grapho-phonemic correspondences [7] are mostly unambiguous from grapheme to phoneme, i.e. the pronunciation of written text is predictable to a high degree. However, as a result of historical spelling some phonemes correspond to more than one grapheme (e.g. /i/ may be represented by six different graphemes: ⟨ι⟩, ⟨η⟩, ⟨υ⟩, ⟨ει⟩, ⟨οι⟩, ⟨υι⟩), hence the spelling of speech is not predictable without dictionary. Consequently, text-to-speech or pronunciation generation, respectively is less problematic than ASR.

Two pronunciation variant dictionaries have been developed using a grapheme-to-phoneme conversion system implemented as a perl script of up to 70 rules with a few exceptions:

- structure words like τον, την (male and female definite article in accusative) and very frequent monosyllabic words were transcribed manually because of their manifold phonetical realisations
- the ⟨γγ⟩-digraph resulting from ‘learned’ formations of the prefixes {εν-, συν-} and stems with initial /γ/, e.g. έγγραφο “document”, is phonetically transcribed as [ŋɣ] (in contrast to [ŋg] as usual)
- company or product names and acronyms written in latin characters (e.g. BBC, Unesco, Löwenbräu) also had to be transcribed manually

For each dictionary, distinct acoustic models have been trained and used during recognition.

3.1. Allophonic phone set (allo)

The first variant lexicon makes use of the complete Greek phone inventory: 26 consonants (except for affricates like /tʃ/, /dʒ/, which were separated as /t/+/s/, /d/+/z/, respectively); 5 vowels, the non-syllabic /i/ and one additional phone for every

stressed vowel. In summary, this results in 37 plus 4 artificial phones (SILence, BReaTh, LIPsmack, GaRBagE) required for acoustic training.

Aside from the phonological processes described above the following phenomena were found to be relevant for phonetic transcription:

- pronunciations of the consonantal digraphs ⟨μπ⟩, ⟨ντ⟩, ⟨γκ/γγ⟩ within words vary between [b], [d], [g] and [mb], [nd], [ŋg] (not at word beginnings) due to regional, stylistic, and individual differences
- digraphs ⟨αυ⟩, ⟨ευ⟩, ⟨ηυ⟩ are pronounced as [af], [ef], [if] before voiceless consonants and as [av], [ev], [iv] before voiced consonants or vowels
- within pronunciations of the digraphs ⟨αύ⟩, ⟨εύ⟩, ⟨ηύ⟩, the vowel has to be stressed, although for reasons of orthography the written accent is put on the consonantal component

Obedying all (approx. 70) rules leads to an average number of 1.97 pronunciations per lexicon entry.

3.2. Reduced phone set (red.)

In a second attempt to construct a recognition dictionary a very simple letter-to-sound mapping has been applied, i.e. phones without immediate graphemic correspondence were neglected. The idea behind was to let the acoustic models learn allophonic variations. Thus, the palatal allophones, the voiced plosives, the velar as well as the labiodental nasal [ŋ] (an allophone of /m/) are ignored, but all 5 vowels and their stressed variants are included. This leaves us with a reduced phone set of 26 phones plus 4 artificial ones. Exceptions to the rules given above (see Section 3) but no further phonological processes or other phenomena were considered for phonetic transcription. The average number of pronunciations per lexicon entry becomes 1.003.

4. Experimental setup

Experiments were carried out using audio recordings (mono, 16kHz sampling rate, 16 bit resolution) of news shows broadcasted in summer 2001 and spring 2002 via the Greek satellite-TV channel EPT “ERT”. Transcription into text as well as XML-annotation (timing, speaker turns and names, non-speech utterances, etc.) of the collected audio data was done manually at ILSP² by means of the Transcriber tool [8]. Recorded speech data were divided into a training set of 36^h05^{min} and a disjoint development test set of 1^h35^{min}.

Acoustic models are context-dependent triphone and quin- phone models derived from mel-frequency cepstra extracted from the audio. Several normalisation and adaptation techniques like cepstral mean subtraction are applied on a per utterance base. Each phone model is a continuous density Hidden Markov Model with either state-clustered or phonetically tied Gaussian mixtures. The recognition engine in use is based on BBN’s Byblos technology [9].

Two text corpora of approximately 2+23 million words were provided by ILSP and had to undergo several preprocessing steps (courtesy of ILSP) in order to obtain clean and convenient data for language modelling. A total number of 25M words produces an exhaustive word list of about 350k different lexical terms of which 200k occur more than one time³.

²Institute for Language and Speech Processing (<http://www.ilsp.gr>)

³For comparison: the text corpus employed for the French CIM-

4.1. Phone set experiments

For phone set experiments data subsets of $4^h 49^{min}$ and 49^{min} for training and testing, respectively have been utilised. Table 1 characterises five (slim) ASR systems with respect to augmenting lexicon and language model (LM) size. Lexical coverages⁴ increase with lexicon size and hit 100% for the *Ooov* system (all test set words included in the dictionary) and the *unfair* case (*Ooov* plus incorporating the test set into the LM). 3-gram LM-perplexities p_3 increase from *mini* over *base* to *Ooov* as more and more infrequent words are included in the lexicon. Accordingly p_3 drops dramatically for the *unfair* system. Out-of-vocabulary (oov) words were discarded for p_3 -computations.

Table 1: Weighted and unweighted lexical coverages C_w, C_u and 3-gram perplexity p_3 for different ASR systems with varying lexicon sizes (lex.) and language models (LM).

	lex.	LM	C_w (%)	C_u (%)	p_3
mini	11304	2M	88.0	66.9	407.9
base	20478	2M	92.6	78.8	437.3
Ooov	21070	2M	100.0	100.0	730.6
unfair	21070	2M	100.0	100.0	4.8
ext.	20478	25M	92.6	78.8	393.2

All the five systems were trained applying three different phone sets, see Table 2 for performance figures. Doubling the dictionary size and increasing the LM data by a factor of approx. 10, will each yield about 4% improvement in WERs. *Ooov* switches off oov-effects resulting in 8% gain compared to *base*. In the *unfair* case we examined just the acoustic component, WER drops to half the *Ooov* value, insinuating the importance of the LM. Only small differences in recognition rates were measured for the corresponding experiments with slightly better results for the *ILSP* phone set. Models originating from the reduced phone set seem to cover phonetic variabilities almost as good as the *ILSP* set, whereas the allophonic set inherently seems to be too ambiguous for that small amount of training data. The *ILSP* set constitutes an effective compromise between number of phones and considering phonological (and other) processes (same number of phones as *allo* and as much pronunciations per lexical entry as *red.*).

Table 2: Word Error Rates (WER) of the systems presented in Table 1 using the *ILSP* (courtesy of *ILSP*), the reduced, and the allophonic phone set.

phone set	ILSP	red.	allo
# of phones	41	31	41
	WER(%)		
mini	50.2	52.3	52.5
base	46.7	47.6	47.2
Ooov	38.6	38.3	39.4
unfair	17.3	18.5	18.4
ext.	42.2	43.0	44.3

The most frequent types of errors are insertions and deletions of common, poorly articulated, short words like negative

WOS system [10] consists of 116M words of which 488k are unique and 286k occur more than once.

⁴The unweighted coverage C_u , as opposed to the weighted C_w , counts every oov-word uniquely independent of its multiplicity.

and modal particles, articles, prepositions, and conjunctions. The introduction of word insertion penalties would at least partially overcome these difficulties. Another source of error is provoked by homophonies of word transitions within different word sequences, which cause wrong word boundary settings, e.g. note the displacement of initial [s] in the REFERENCE (Σ, σ) to final [s] in the HYPOTHESIS (ζ):

REF: ... στη Λεωφόρο Σπάτων στη ...
HYP: ... στη Λεωφόρος πάντως τη ...

A well endowed language model seems to be the only way out in this case, provided corresponding words are not oov.

4.2. Lexical and language modelling

As illustrated in Section 2.3, the inflectional degree of Modern Greek poses extraordinary standards to the recognition dictionary as well as the language model. Corresponding investigations were performed using ASR systems trained and tested on the respective total data sets (Section 4). Lexica are assembled by taking all words from the audio transcripts as a basis⁵, and extending it by those words of the text corpora with frequencies higher than a given cut-off or threshold, cf. Figure 1. This is

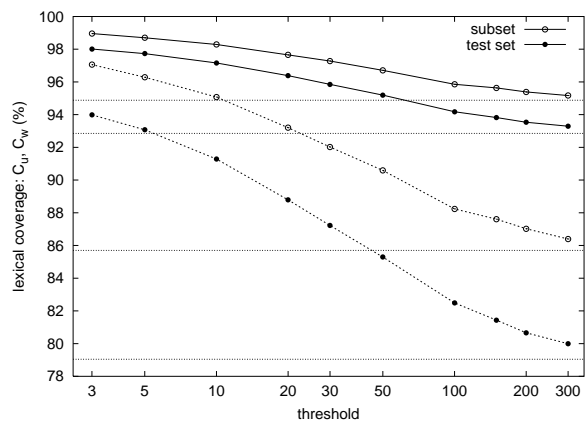
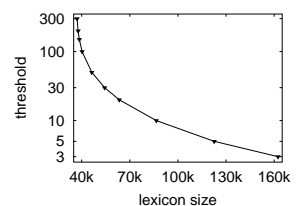


Figure 1: Weighted C_w (solid lines) and unweighted lexical coverages C_u (dashed lines) obtained by the set of words with occurrences of more than a minimum-threshold in the text corpus. Offsets, indicated as straight lines, are due to word contingents from the audio transcripts.

why coverage figures approach a saturation level for increasing cut-off (the straight lines in Figure 1). No more than a threshold value of 3, i.e. including words with rather small unigram probability, lead to coverages generally reported for recognition dictionaries of comparable utility [10].

Now what are the implications on the lexicon size? Figure 2 depicts coverages as a function of the number of lexical entries.

As one can read off the diagram alongside, coverages due to a cut-off of 3 pertain to a dictionary of more than 160k terms. However, in Figure 3, displaying word error rates versus lexicon size, we observe stagnating



⁵The audio vocabulary is supposed to be more similar to the ASR's actual operational area than that of the text corpora.

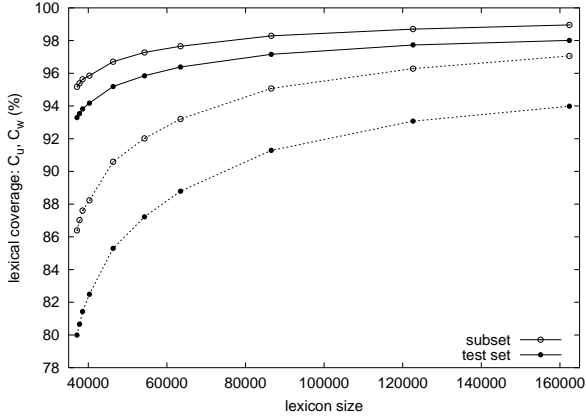


Figure 2: Weighted C_w (solid lines) and unweighted lexical coverages C_w (dashed lines) for increasing size of recognition dictionaries.

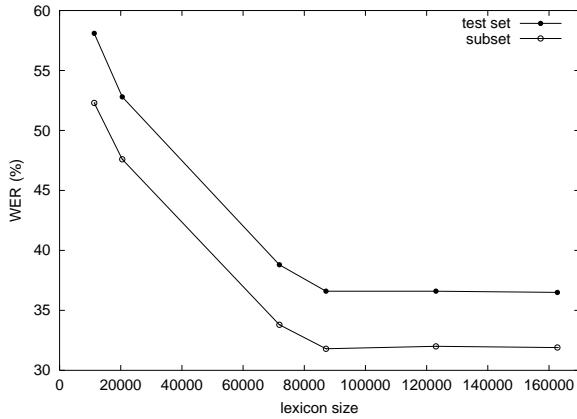


Figure 3: Word Error rate against recognition lexicon size for ASR systems trained by using the total audio data and the reduced phone set.

values for sizes greater than approx. 90k. Additional words of low frequency don't reduce word error rates further as support by the language model collapses due to non-occurrence of corresponding n -grams ($n = 3$ in our case). Language modelling with

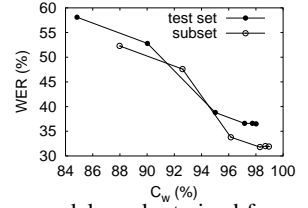
$$P(S) \approx \prod_{i=1}^n P(w_i | w_{i-n+1}, \dots, w_{i-2}, w_{i-1}) \quad (1)$$

denoting the probability of a sequence S of words w_i , has to be performed on considerably larger text corpora in order to get sufficiently accurate n -gram probabilities $P(w_i | w_{i-n+1}, \dots, w_{i-1})$. Even backing-off from tri-gram to bi-gram and uni-gram probabilities seems to be futile here.

5. Conclusions & Perspectives

We presented a Modern Greek grapheme-to-phoneme conversion system and examined its impact on performance for various ASR systems. In view of the complex morphological structure of Modern Greek several experiments were accomplished employing recognition dictionaries of different extent.

Word Error rates range between 30% and 40% for lexical coverages of greater than 95%. A rough estimation reveals that a Greek lexicon would require a multiplicity of entries more, than one of comparable utility



for English, provided the language model can be trained from sufficiently large corpora to avoid the n -gram sparseness problem. Concurrent ASR systems for inflectional languages, e.g. Czech [11] try to solve the problem of enormous vocabulary growth by performing automatic stemming and sophisticated morpheme-based language modelling. These techniques require grammatically tagged corpora and a morphological lexicon. However, as argued in Section 2, morphological decomposition is extremely non-systematic for Modern Greek and thus difficult to implement by means of a rule-based stemming software. Therefore, for the time being, we stick to full form word lexica and expect improved word error rates by incorporating more and more language model data.

6. References

- [1] <http://www.xanthi.ilsp.gr/cimwos/>
- [2] Holton, D., Mackridge, P., Philippaki-Warbuton, I., *Greek. A Comprehensive Grammar of the Modern Language*, London/New York, 1997. Mackridge, P., *The Modern Greek Language*, Oxford, 1985.
- [3] Setatos, M., *Phonological Problems of Modern Greek Koine*, Thessaloniki, 1969. Setatos, M., *Phōnologia tēs Koinēs Neohellēnikēs*, Athens, 1974. Ruge, H., *Nygrekisk Fonetik*, Stockholm, ³1979.
- [4] Tonnet, H., *Manuel d'accentuation grecque moderne (démotique)*, Paris, 1984.
- [5] Kinne, D., *Diglossie und Bistrukturalität in den Texten der griechischen Verfassung von 1975 und 1986*, MA Thesis, FASK Germersheim, Univ. of Mainz, 2001.
- [6] Eleftheriades, O., *Modern Greek Word Formation*, Minneapolis, 1993.
- [7] Katsikas, S., "Probleme der neugriechischen Graphematik aus der Perspektive des Fremdsprachenlernens", in Eichner, H., et al. (eds): *Sprachnormung und Sprachplanung*, 419–474, Wien, ²1997.
- [8] Barras, C., Geoffrois, E., Wu, Z., and Liberman M., "Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech", 1st International Conference on LREC, 1373–1376, 1998.
- [9] Makhoul, J., Kubala, F., Leek, T., Liu, D., Nguyen, L., Schwartz, R., and Srivastava, A., "Speech and language technologies for audio indexing", Proc. of the IEEE, Vol. 88: 1338–1353, 2000.
- [10] Hecht, R., "French Broadcast News Transcription", these Proceedings.
- [11] Byrne, W., Hajič, J., Ircing, P., Jelinek, F., Khudanpur, S., Krbeč, P., Psutka, J., "On Large Vocabulary Continuous Speech Recognition of Highly Inflectional Language - Czech", Proc. of Eurospeech 2001, Vol. 1: p 487, 2001.