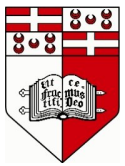
Text to Speech Technologies for Mobile Telephony Services

Paulseph-John Farrugia

Department of Computer Science
University of Malta



March 2006

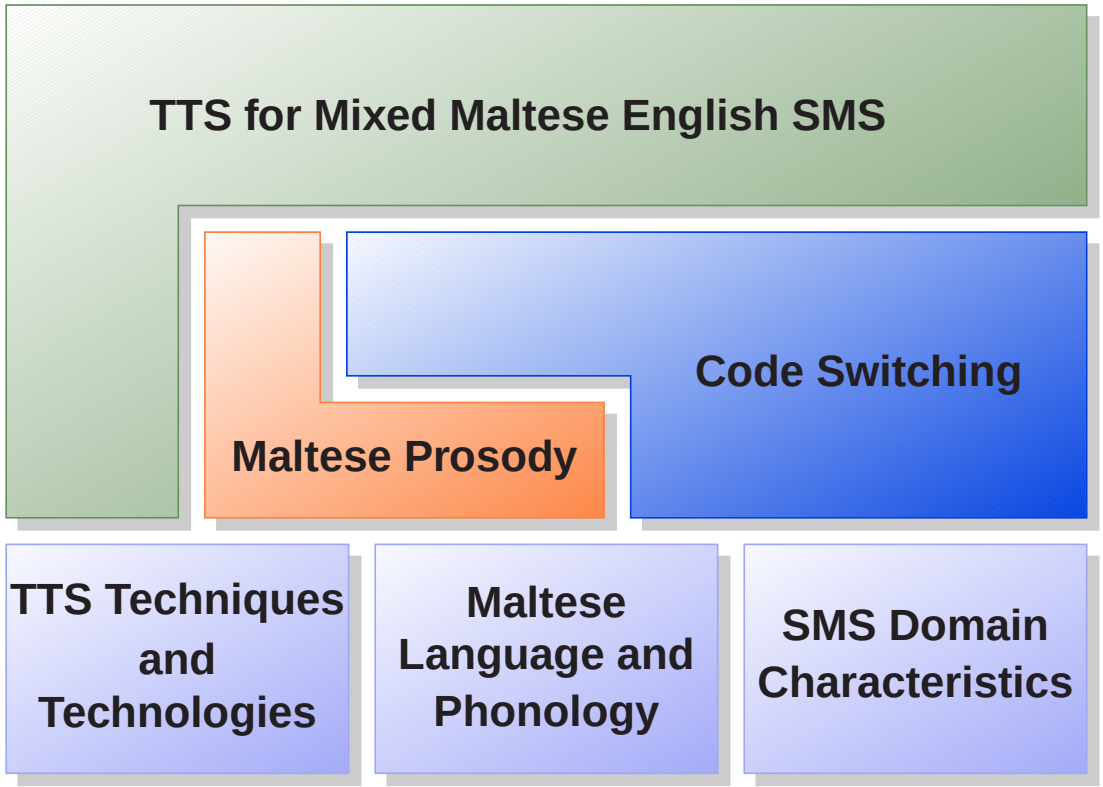
Outline

1. Introduction
   - Aims and Objectives
   - Research Organisation
2. Background Research
   - Text to Speech
   - Maltese
   - SMS Messaging
3. Project Development
   - Code Switching
   - Maltese Prosody
   - Text to Speech Implementation

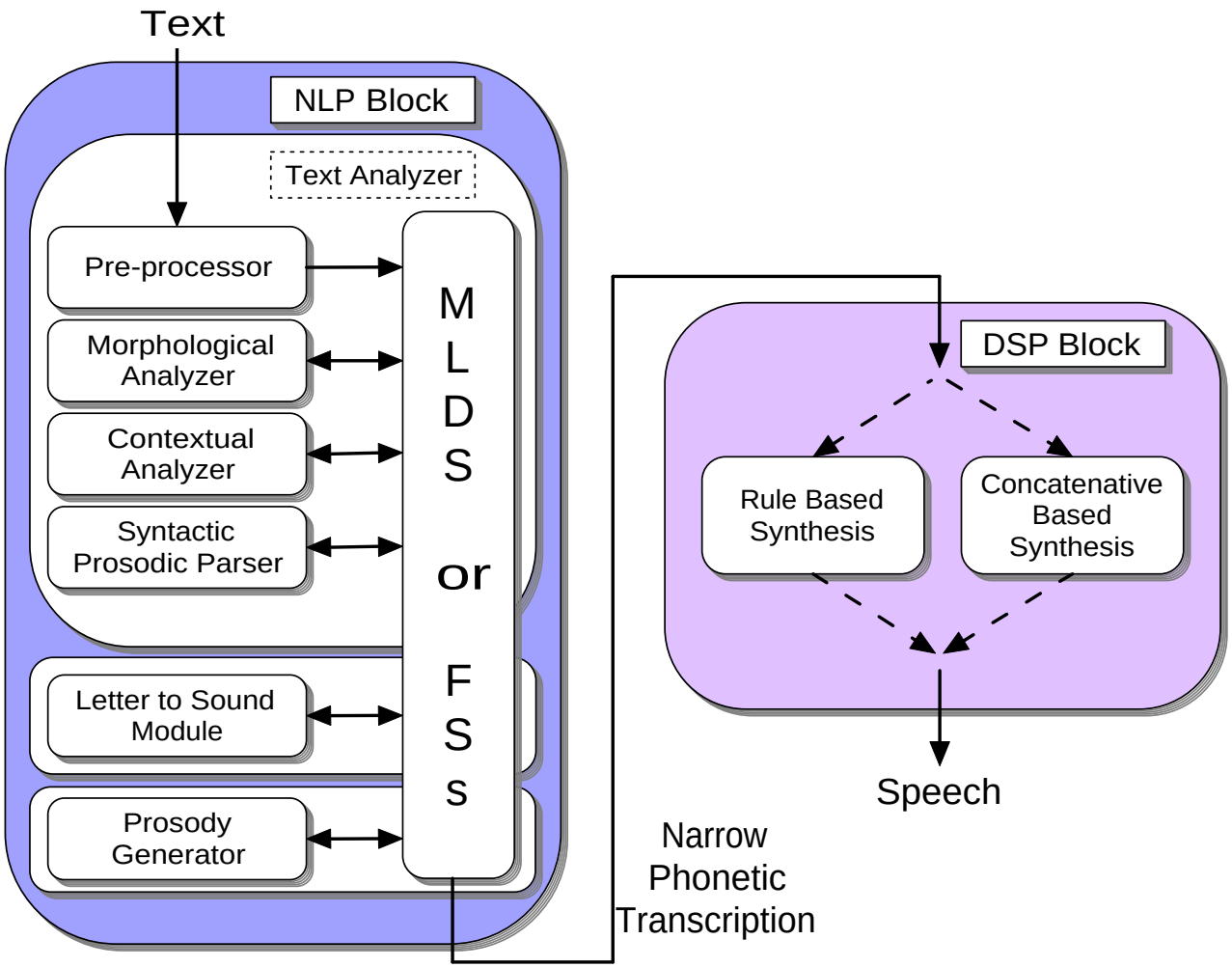
Research Aims

- The project's overall aim was that of implementing a TTS system for SMS messages within the local context, thus enabling unified messaging possibilities.
- The project also provided a canvas for focusing on two main research objectives:
  - Handling the code switching between Maltese and English that naturally occurs in local written (and verbal) communication.
  - Investigating prosodic modelling for Maltese with the intent of applying this within the synthesis process and hence achieve more natural results.

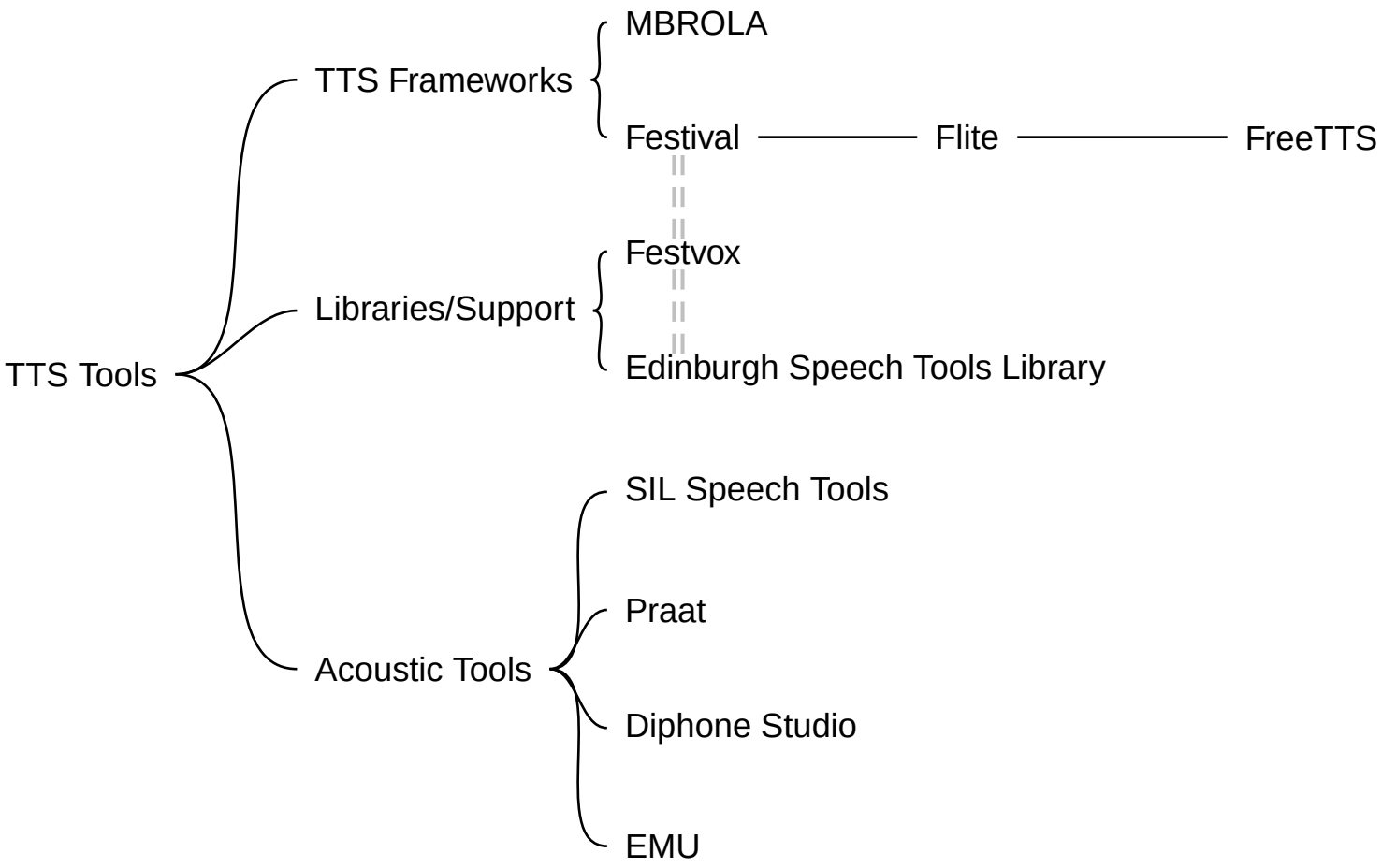
Research Organisation



Text to Speech System General Architecture

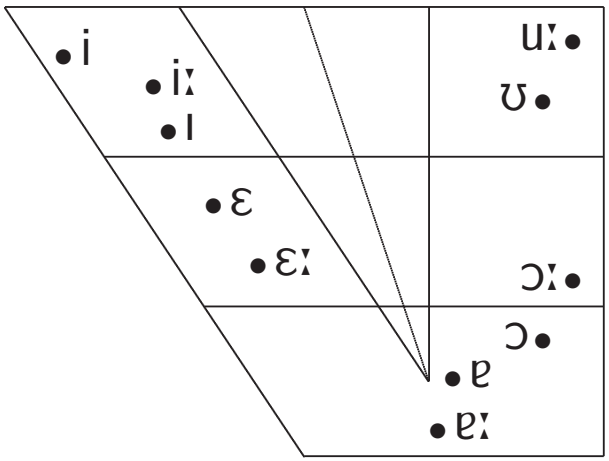


Tools and Resources



Maltese Phonemic Inventory

	Bilabial	Labio-Dental	Alveolar	Post-Alveolar	Palatal	Velar	Glottal
Plosive	p b		t d			k g	ʔ
Nasal	m		n				
Fricative		f v	s z	ʃ			h
Affricate			ts dz	tʃ dʒ			
Retroflex			r				
Lateral Approximant			l				
Approximant	w				j		



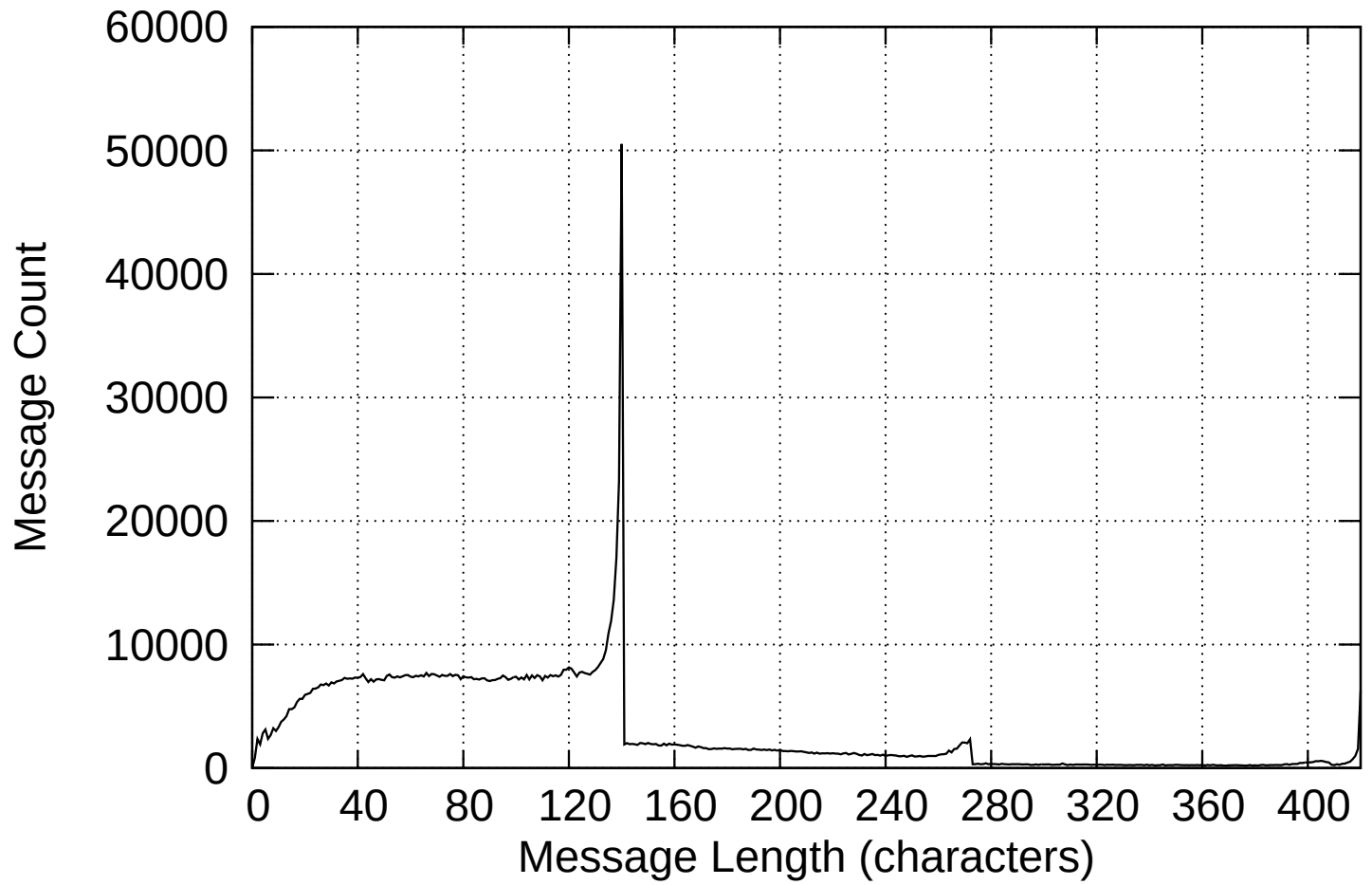
Orthographic to Phonemic Relationship

- The relationship between orthography and phonetics in Maltese is relatively straightforward but subject to a few exceptions, mostly limited to a small set of noun/verb ambiguities:
  - E.g. ‘ħajjat’ → /'haj-jat/ (English *he sewed*) or /haj-'ja:t/ (English *tailor*)
- Other anomalies also exist:
  - E.g. ‘tlieta’ → /klɪ:tɐ/ rather than /tlɪ:tɐ/.
- Computationally, the relationship can be, for the larger part, captured using prioritised, two-level morphological rules of the form $a \rightarrow b/c\_d$ meaning “rewrite orthographic *a* as phonemic *b* when it occurs between orthographic *c* and *d*”.

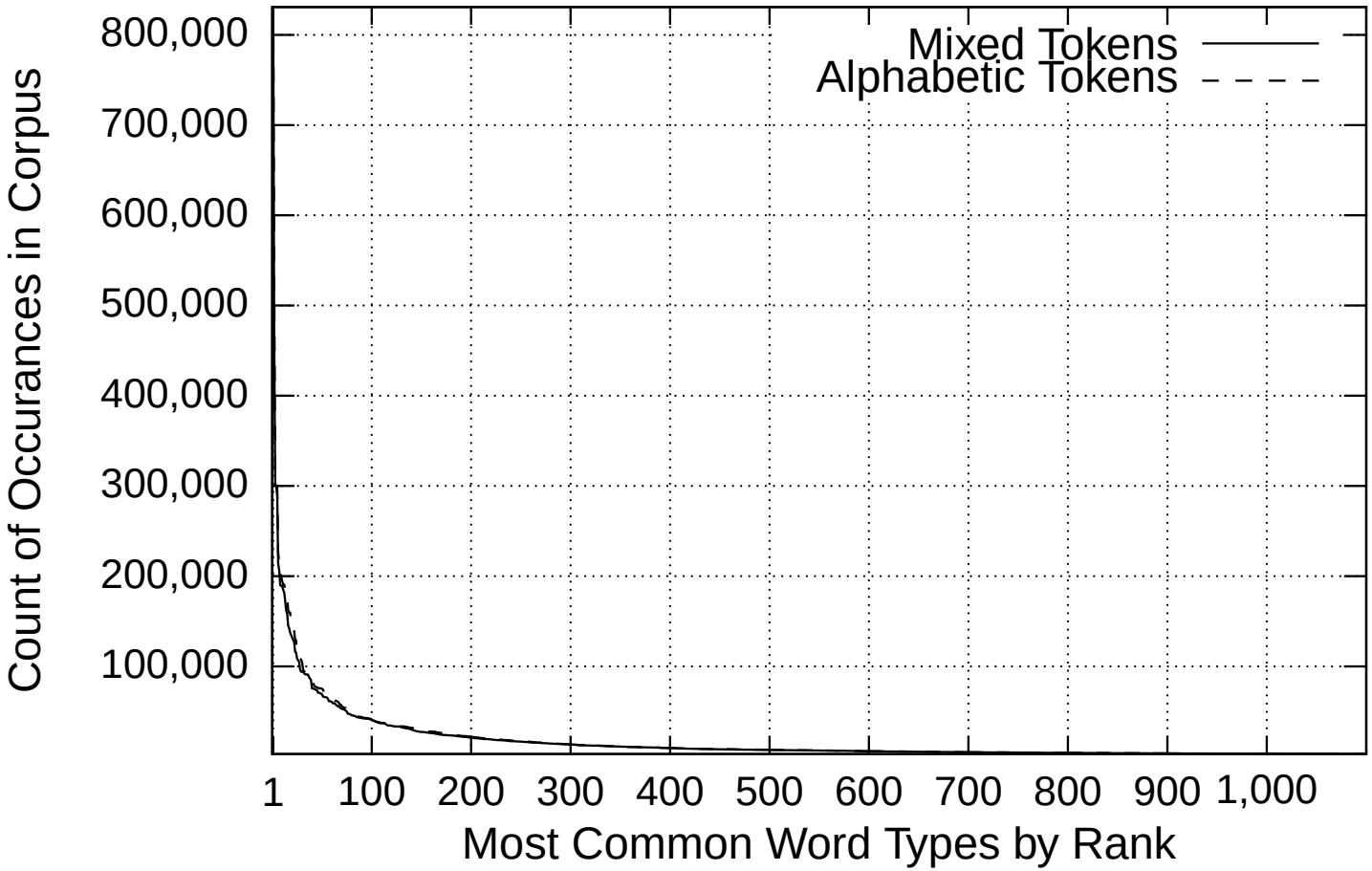
SMS Technical Overview

- A single SMS message can be up to 160 characters long, each character having 7 bits.
- The protocol also supports message concatenation, as well as Unicode and binary data encodings.
- SMS is a store and forward service. Messages are are always routed through one or more SMS Centres (SMSCs).
- SMS messages can be sent and received simultaneously with GSM voice, data and fax calls.
- SMS limitations have given rise to a new manner of writing, often linguistically abrupt and with various abbreviations or smileys.

SMS Data Set Length Histogram



Word Count Frequency



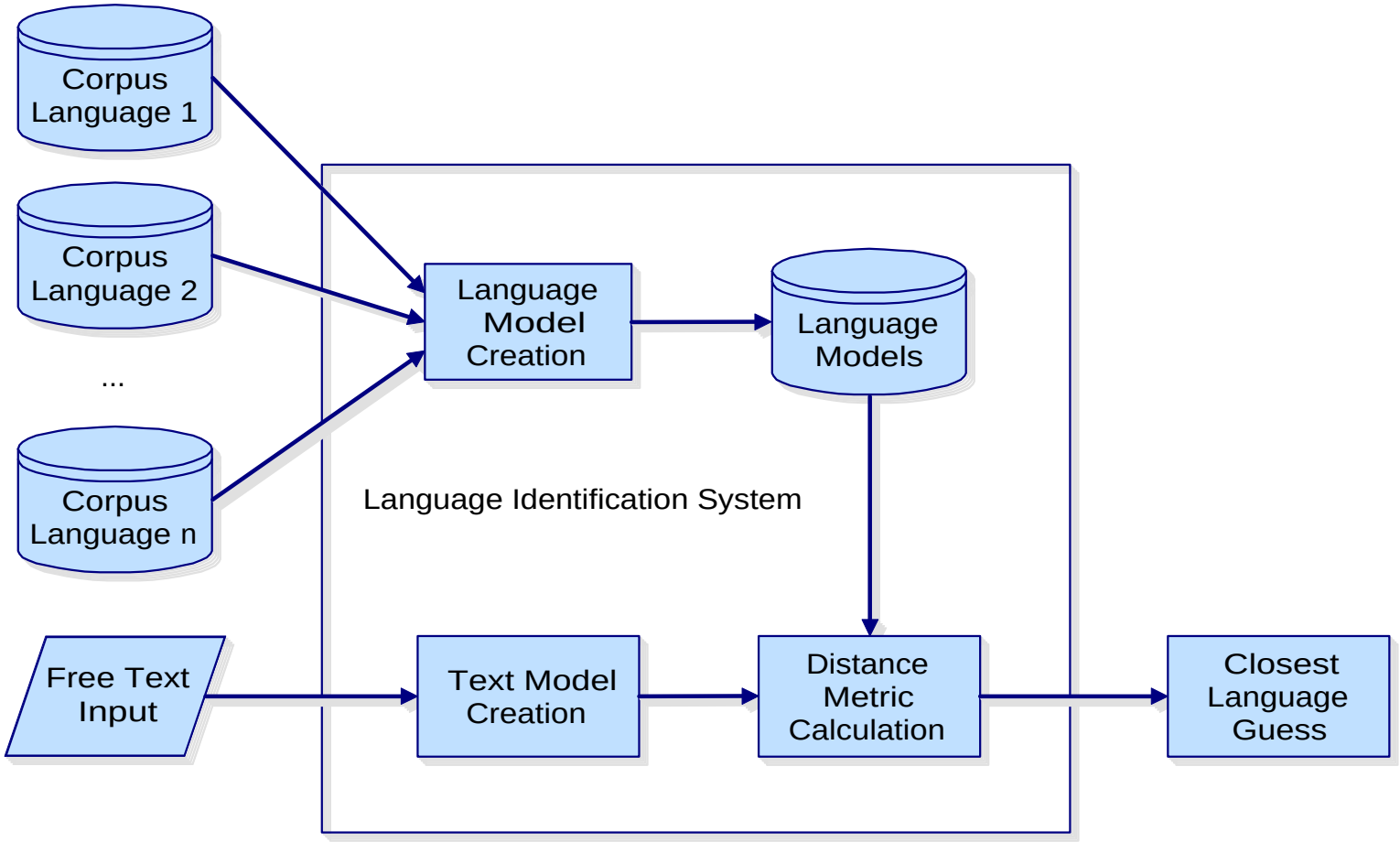
Mixed Maltese English

- As Maltese, we exhibit a tendency to code switch for various reasons.
- Assimilation of English in Maltese
  - Orthographic adaptation to reflect phonological rules
    e.g. *cake* → *kejk* (/kɛjk/).
  - Morphological inflection
    e.g. *to stretch* → *tistreċċja* (/tɪstreʧːjɐ/).
  - Partial assimilation: English words appear unmodified, with or without any orthographic indication.
- Implications for processing
  - Lexical, Grammatical, Phonological, Semantic

Code Switching

- Visualising the resulting text as a new language variety.
  - Processing as a monolingual text with classic parsing strategies.
  - Need of combined resources for participating languages.
  - Linguistic interplay and structures not extensively studied.
  - Unknown words and spelling mistakes an issue.
- Disambiguating with stochastic techniques.
  - Use of lookup dictionaries not sufficient.
  - Handling of ambiguous words (e.g. *u* as ‘u’ or ‘you’).
  - Classifying unknown words.

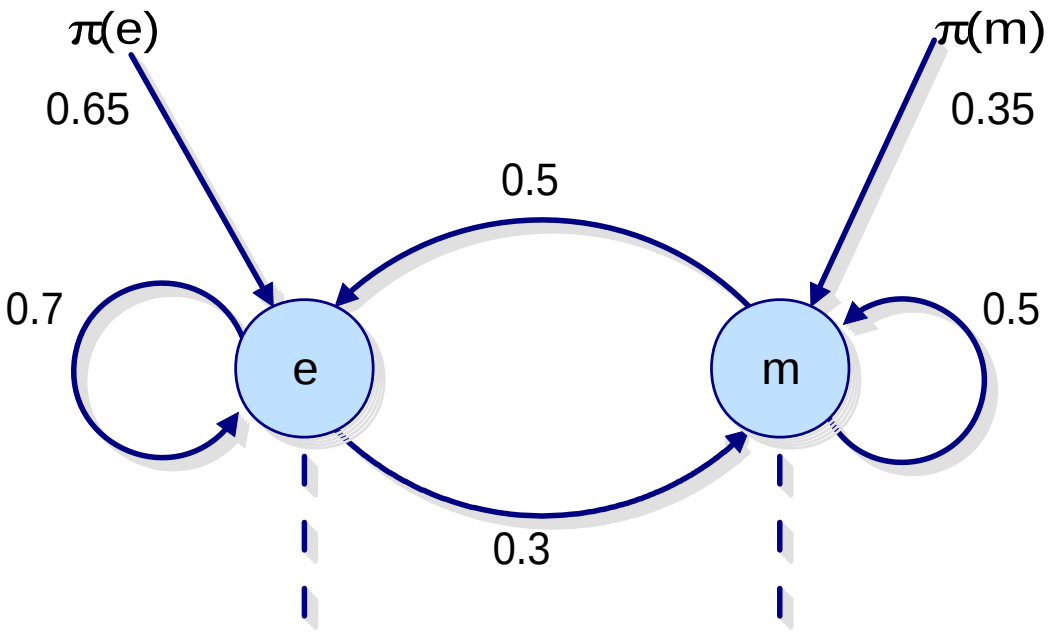
Language Identification General Architecture



HMM Language Tagging

- Code switch parsing as an HMM language tagging exercise.
- Underlying hypothesis - Code switching follows grammatical patterns that can be captured stochastically.
- Need a (manually) tagged corpus to build HMM parameters.
- Viterbi algorithm to determine optimal sequence assignment.

An Example



b(u) = 0.0732	b(u) = 0.0151
b(you) = 0.0546	b(ma) = 0.0312
b(to) = 0.0162	b(qed) = 0.0219
...	...

HMM Parameters

- States represent n-gram language sequences.
- Transitional Probability

$$a_{ij} = P(L_j|L_i) = \frac{count(L_i L_j)}{count(L_i)}$$

- Observation Probability

$$b_i(w) = P(w|L_i) = \frac{count(w{=}w_i)}{count(w_i)}$$

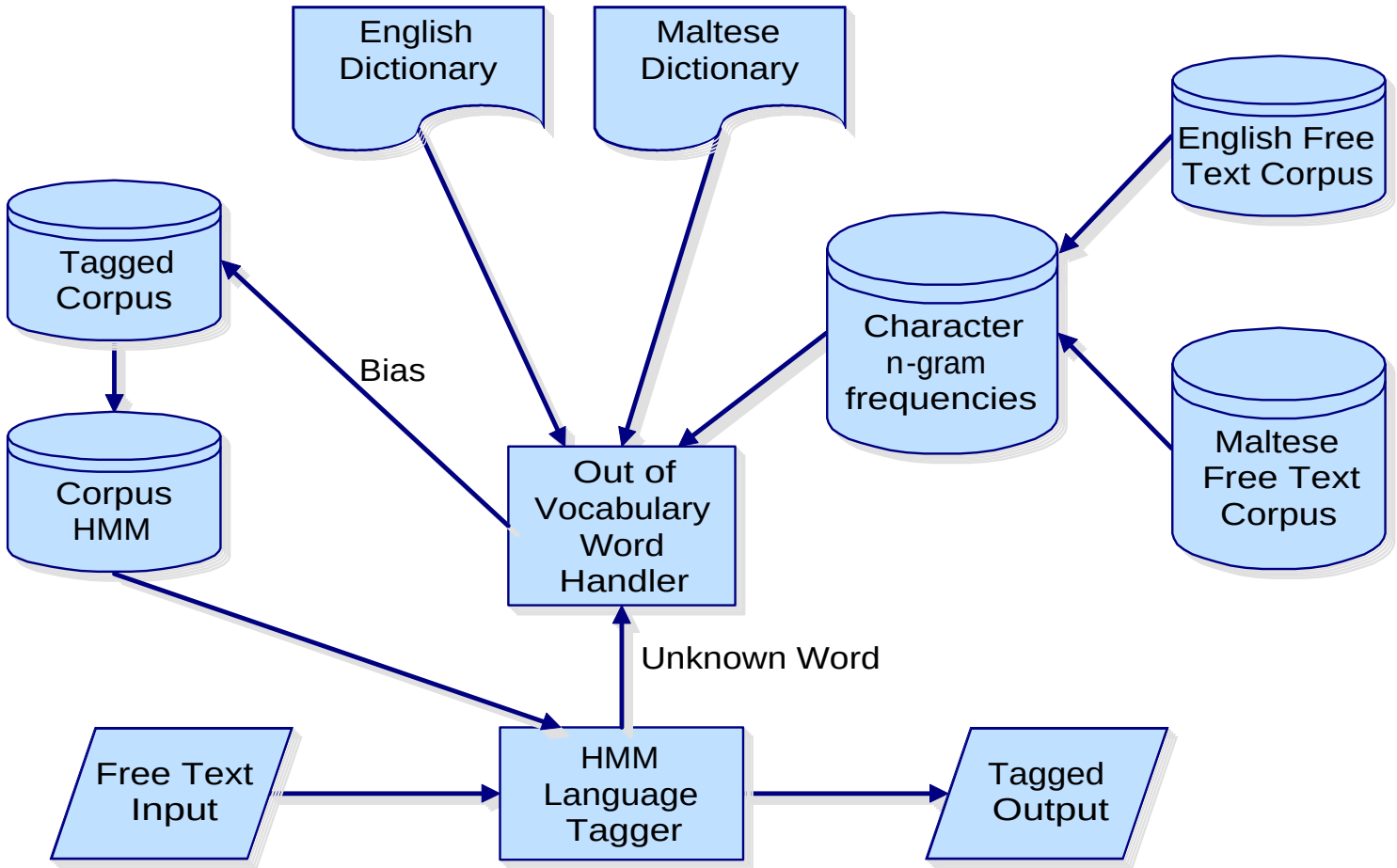
- Initial Probability

$$\pi(L_i) = \frac{count(sample{=}L_i...)}{count(samples)}$$

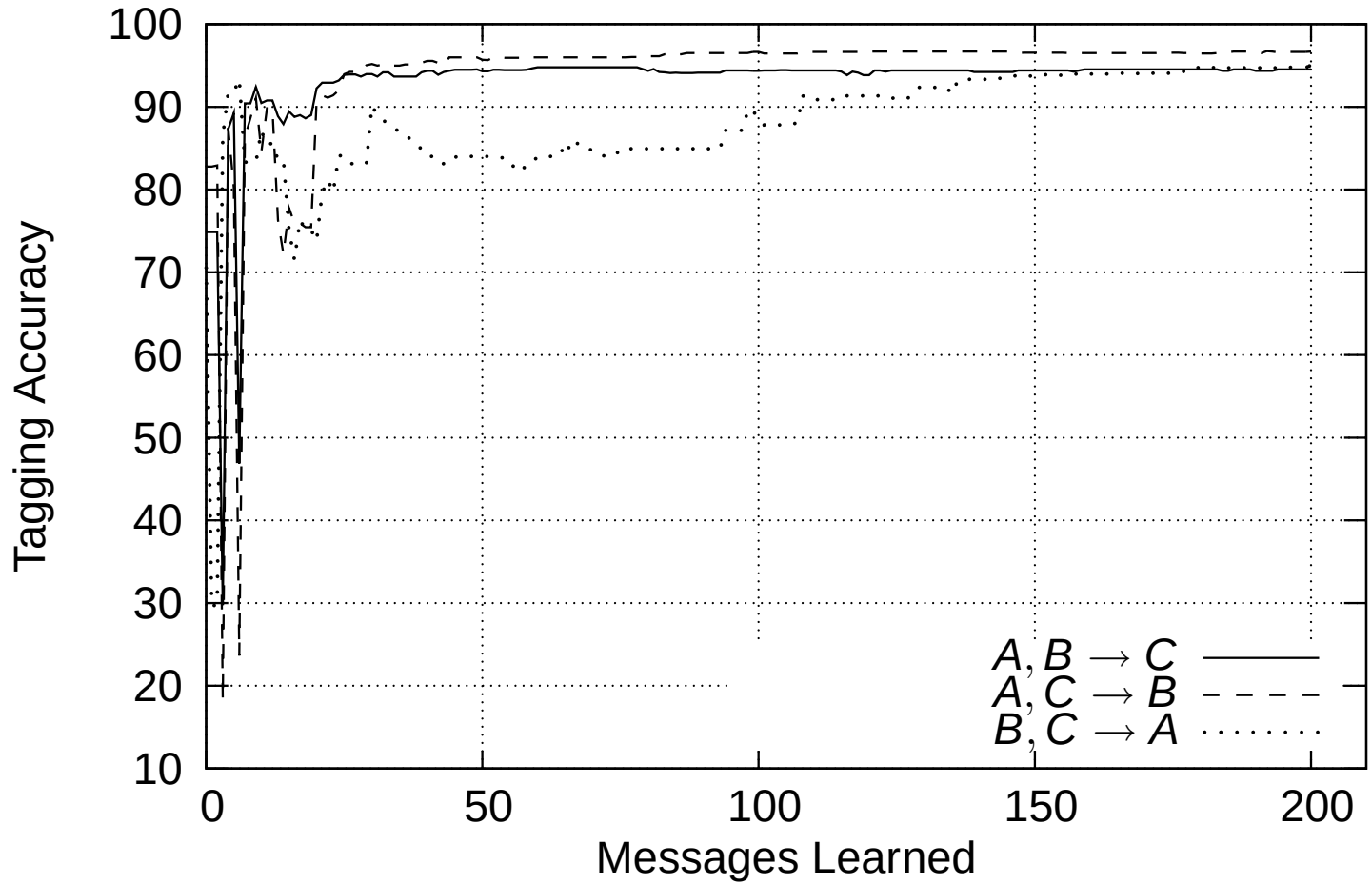
Unknown Words

- Words not found in the tagged corpus will have zero probabilities assigned.
- Smoothing techniques are generally used in this respect.
- In our case, it possible to assign probabilities to non-frequent words in a more meaningful manner.
  - Use of lookup dictionary to bias preference.
  - If dictionary fails, use character n-gram frequencies to make choice.

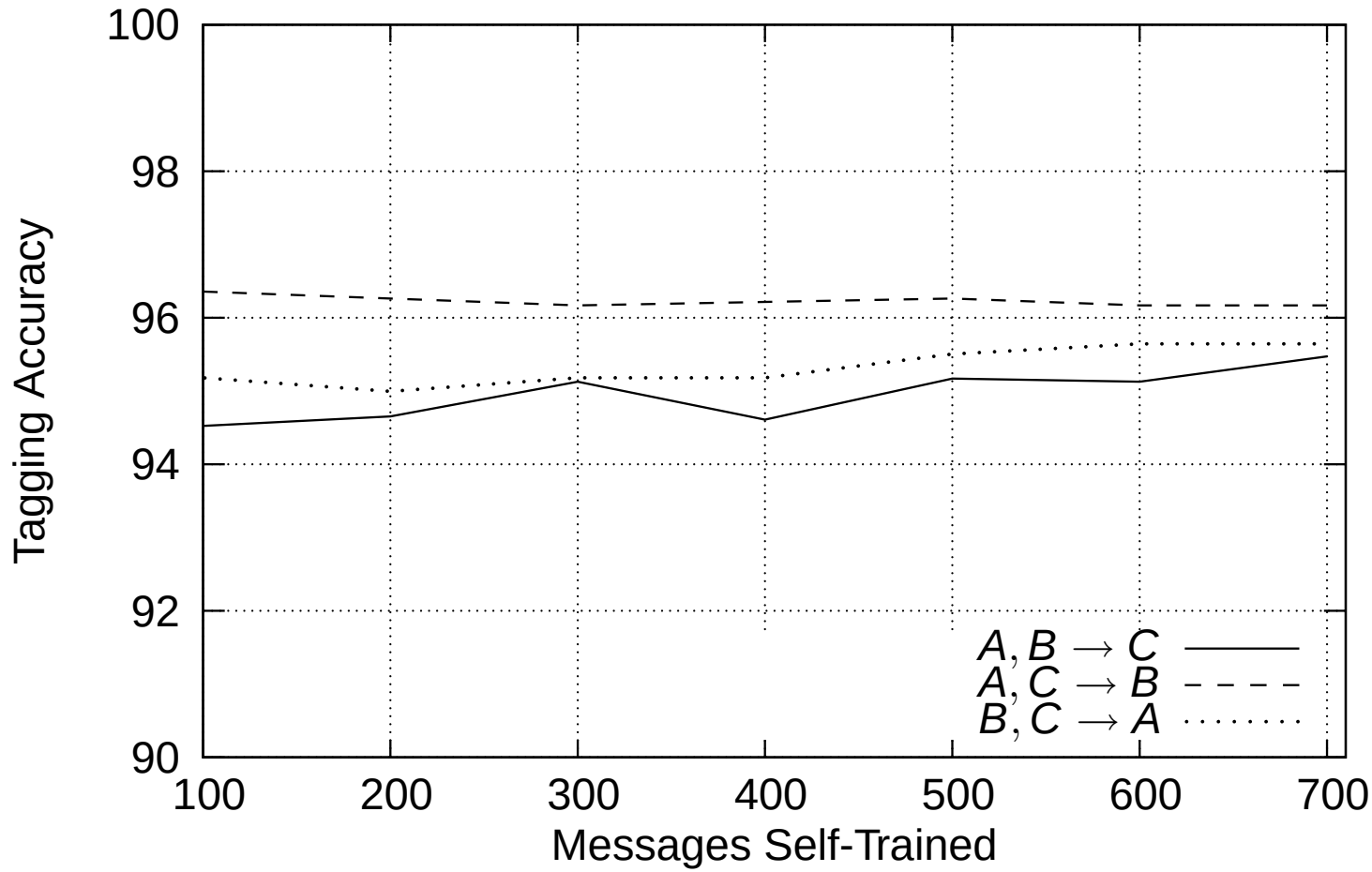
System Architecture



Tagging Improvement



Cross Training



ToBI for Maltese

- A well-recognised framework within linguistic circles, and possibly the most commonly utilised intonational annotation formalism in recent works.
- Its use is prevalent in TTS research.
- Previous linguistic research on Maltese intonation utilisng ToBI-style annotations.
- The formalism is independent of any prosodic assignment technique utilized, thus allowing the exploration of alternative methods within a modular TTS system.

ToBI Development

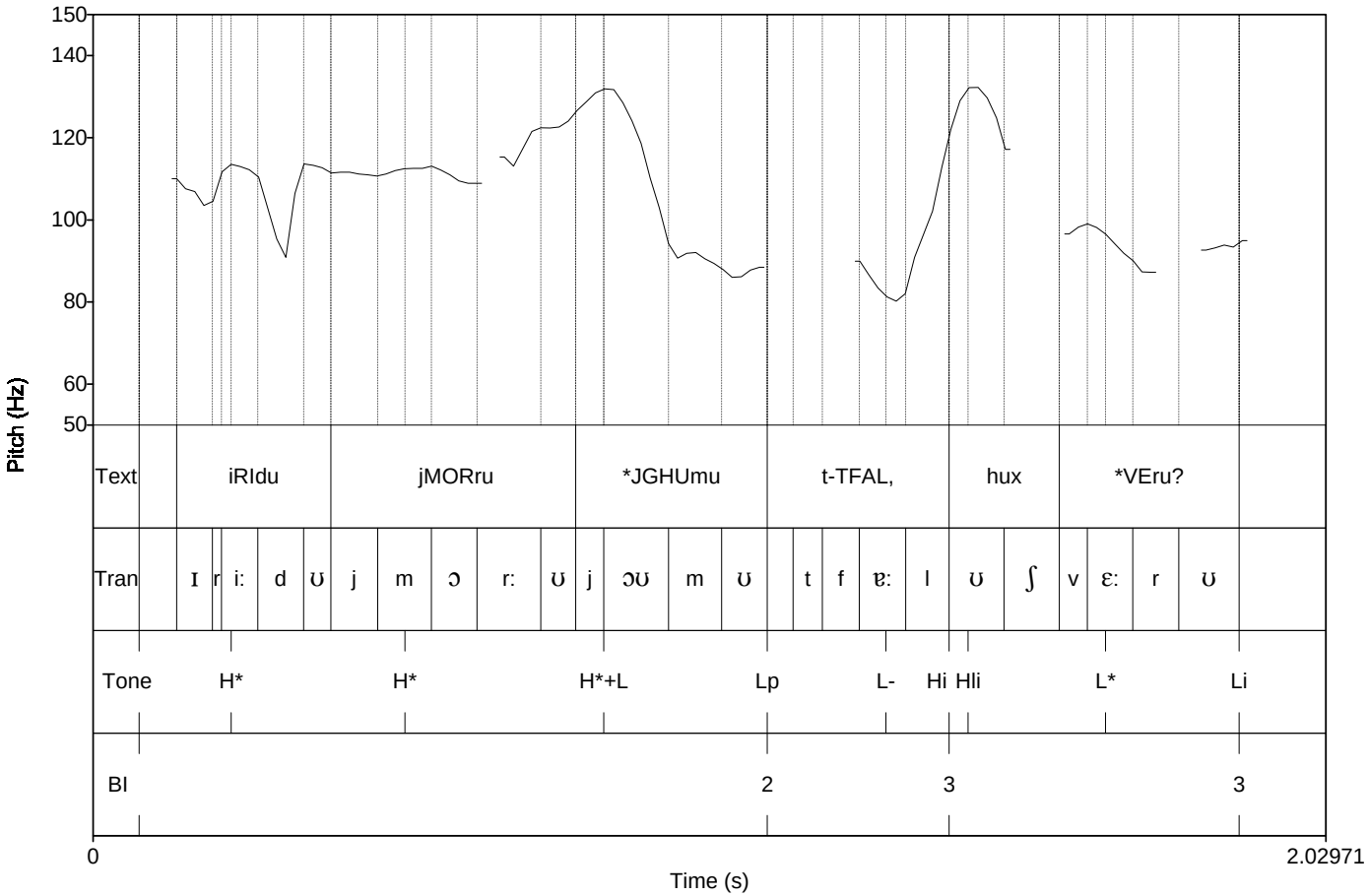
Given the previous phonological research, the remaining requirements for a complete Maltese ToBI system consist of the following:

- The design and recording of a structured speech corpus.
- The formalisation of the intonational descriptions available in previous research in a structured form.
- The establishment of a break index scheme.
- The annotation of the recorded speech database.
- The compilation of annotation guidelines.
- The testing for inter-transcriber reliability.

Designing the Corpus

- The corpus designed was composed of four sections, intended to provide different types of speaking conditions.
  1. The first section consisted of a map task.
  2. The second part of the corpus consisted of a short story to be read out by a single speaker.
  3. The third section of the corpus required a single speaker to read out a sequence of mixed utterances, consisting of statements, imperatives, yes-no questions and tag questions.
  4. The fourth and final section of the corpus consisted again of a paired speaker task, this time comprised of a completely open, but topicalised, conversation.
- Project to be presented at ISCA Speech Prosody 2006.

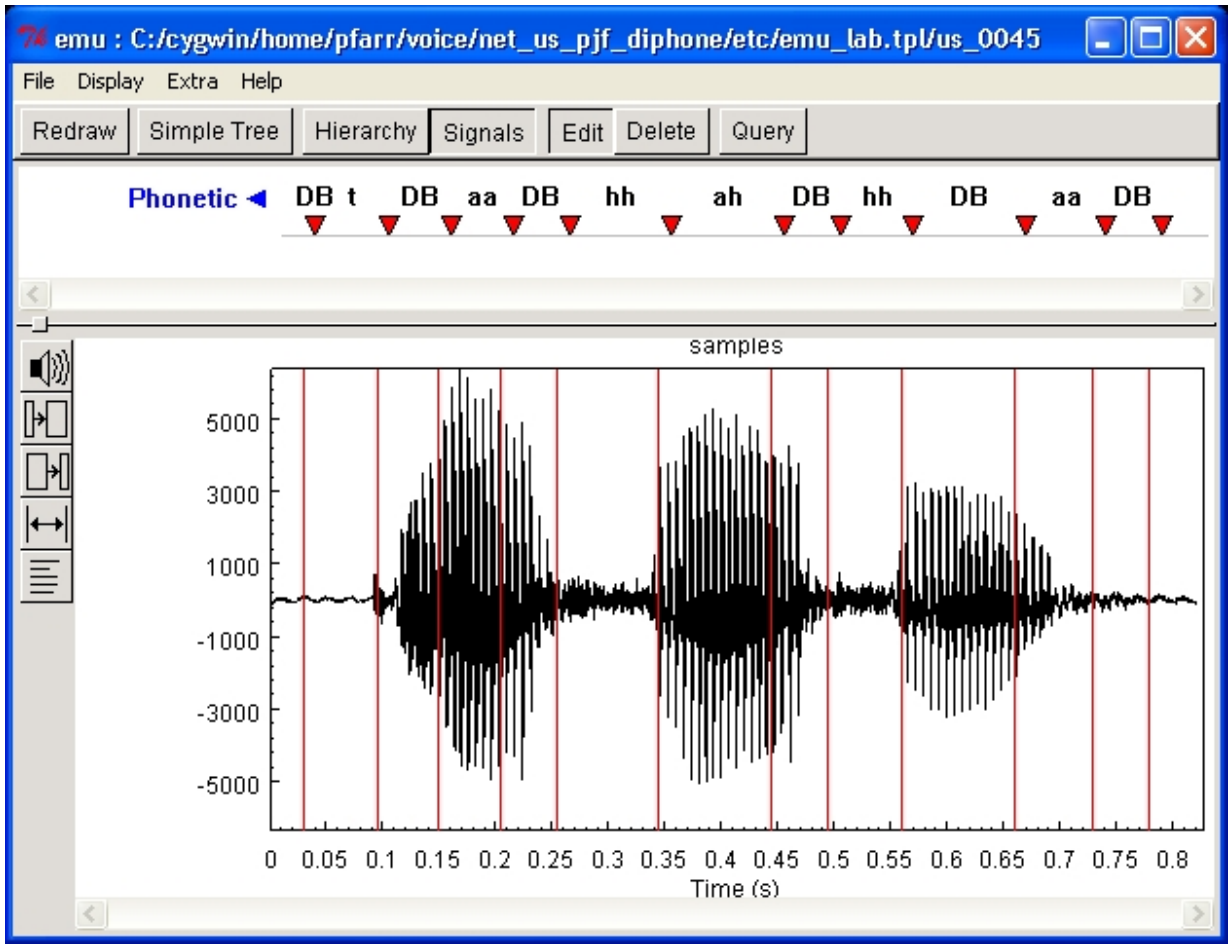
MalToBI Example



Festival TTS System Development

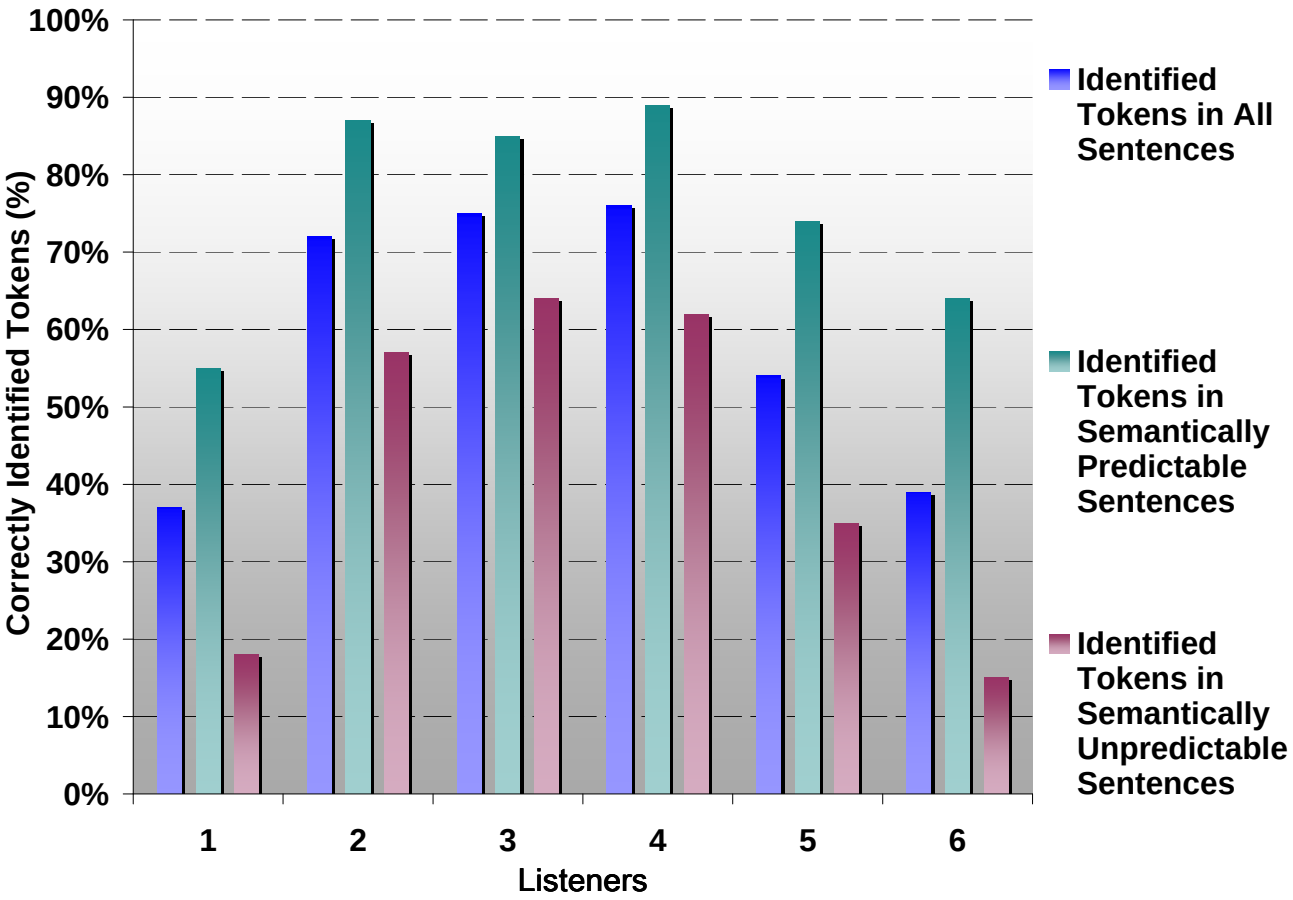
Construct Basic Template Files
⇓
Generate Phoneset Definition
⇓
Generate Diphone Schema File
⇓
Generate Prompts
⇓
Record Speaker
⇓
Label Nonsense Words
⇓
Extract Pitchmarks and LPC Coefficients
⇓
Test Phone Synthesis
⇓
Add Lexicon/LTS Support
⇓
Add Tokenization
⇓
Add Prosody
⇓
Test and Evaluate Voice
⇓
Package for Distribution

Labelling the Carrier Phrases



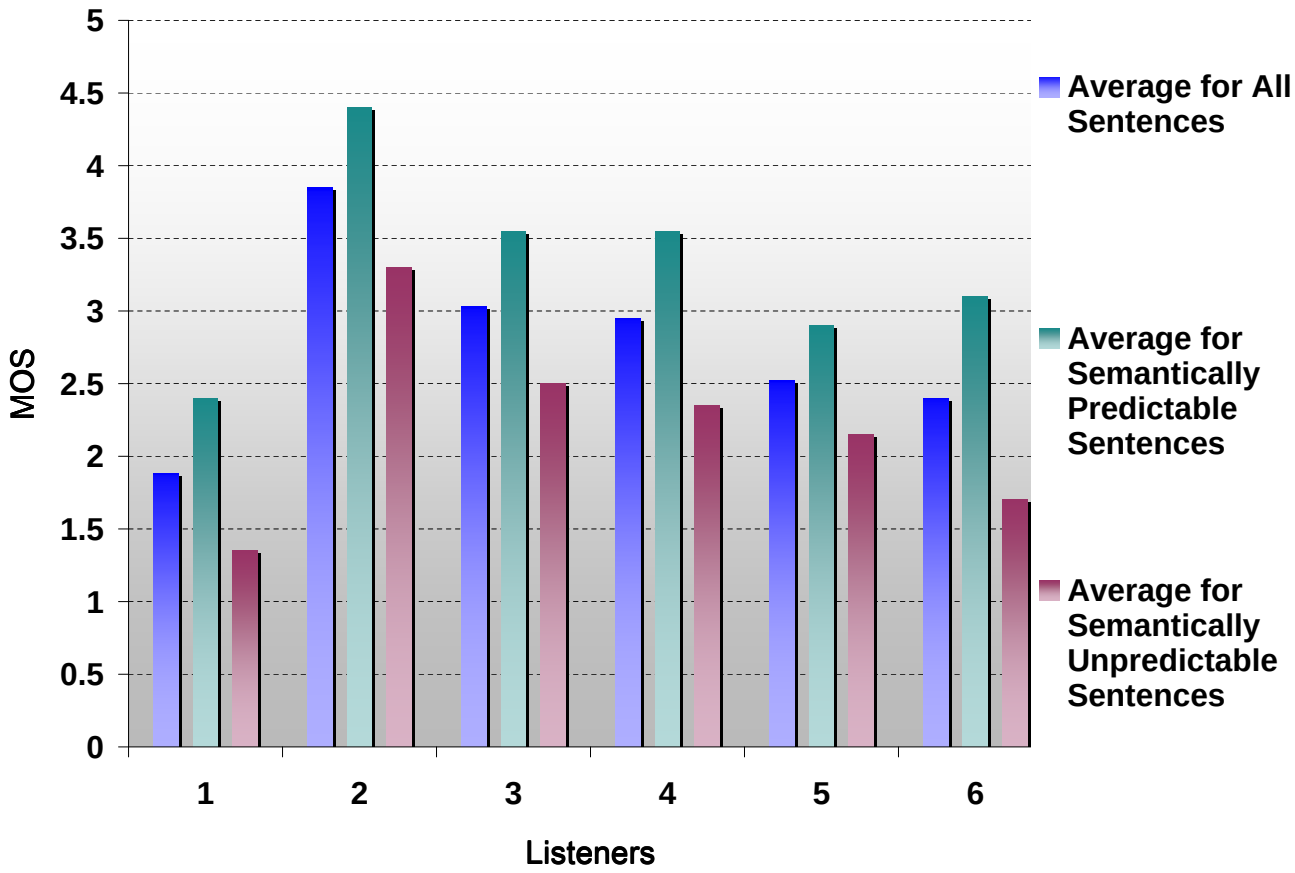
TTS Evaluation

Token Identification

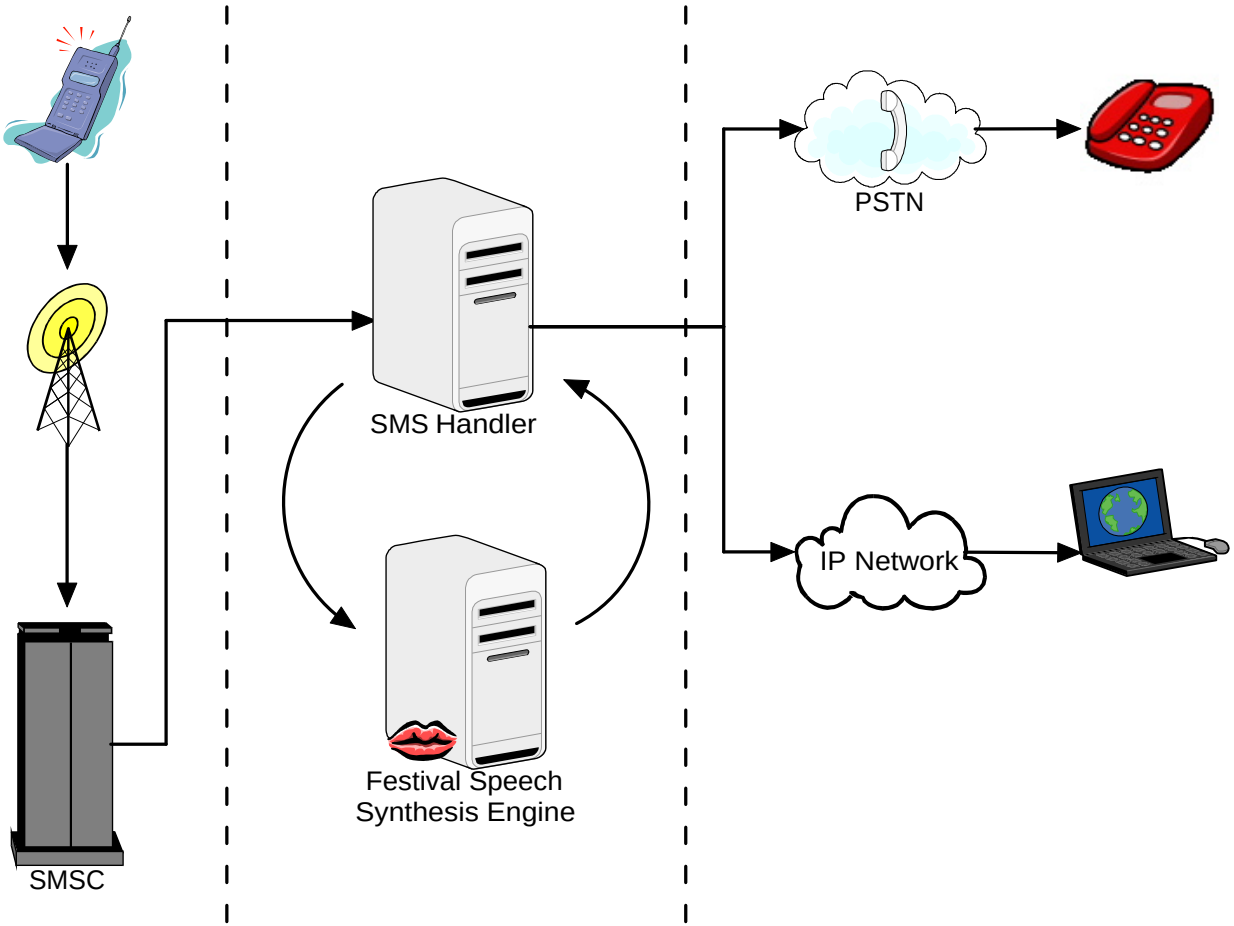


TTS Evaluation

MOS Results



Building An SMS to Speech Gateway



Achievements

- Developed a mechanism for handling code switching in a noisy domain with 95% accuracy.
- Built the case for a prosodic model in the shape of a ToBI annotation system adopted for Maltese.
- Initiated a project aimed to formalize such a system and provide an annotated corpus of native Maltese speech.
- Developed a concatenative-based TTS system for Maltese, also supporting the possibility of code switching with English based upon the language tagger output.
- Developed additional peripheral resources.

Indications for Future Work

- Take into account of how other syntactic features, such as punctuation and non-alphabetic tokens, can be used in the HMM tagging process.
- A pre-processing component that could also be useful in the development of other services, such as an automated mixed language spell checker.
- Further work on the ToBI corpus to serve as a useful resource for both improved TTS results as well as for general prosody research on Maltese.
- Heterophonic homograph and numeric entity resolution.
- Improving voice quality.
- Evaluating LTS accuracy.