COMPUTATIONAL POETRY:

TRANSLATING GREEK AND ITALIAN POPULAR SONGS

Nikitas Tampakis

Advised by: Chrstiane Fellbaum

May 5, 2014

A senior thesis submitted to the Computer Science Department
of Princeton University in partial fulfillment of the requirements
for the degree of Bachelor of Arts.

This thesis represents my own work in accordance with University regulations.

I. Introduction and Motivation

Machine translation (MT) is the use of computer software to translate bodies of texts across languages. State of the art MT systems facilitate global communication and aid in making human languages universally accessible. Google Translate is the de facto platform used for instant machine-based translations of texts. Although the statistical methods used by the system generally produce fairly reliable translation results, they are not designed to handle poetry. Since most automatic translation systems are guided by already existing parallel versions of texts written in multiple languages, they only work well with previously observed language. Poetry is a challenge to translate because it is formed differently from standard language. Observe the following excerpt from the *Axion Esti*, written by Greek Nobel Laureate poet, Odysseas Elytis. The original Greek is accompanied by a Google Translate result as well as a translation by scholar Edmund Keeley. The MT result neither makes sense nor is poetic.

Της δικαιοσύνης ήλιε νοητέ	Justice ilie Sensible	Intelligible sun of justice
και μυρσίνη εσύ δοξαστική	myrtle and you glorifying	and you, glorifying myrtle
μη παρακαλώ σας μη	not please you not	do not, I implore you, do not
λησμονάτε τη χώρα μου	forget my country	forget my country
Greek original	*Google Translate*	*Edmund Keeley translation* **[1]**

Song lyrics, for example, are often more governed by the “grammar” of the musical line than of the language they are written in. Even in just trying to deduce the meaning of a poem written in another language, popular machine translators do not do a particularly good job because they are unlikely to contain reliable data on the unusually-constructed phrases found in poetry. Beyond its lexical meaning, song lyrics also are usually appreciated for a particular rhyme sequence, rhythmic meter, or musical emphases as well.

Since poetic form is a significant component of song lyrics, an ideal song translation would not only make semantic sense, but also have the ability be performed with the original music. This is in fact what happens with songs that get “covered” in other languages. There is a wide degree of variation of writing styles observed across songs, which reflects diversity among musical genres as well as language’s generative nature overall **[2]**. For this reason, there is not a “one size fits all” approach to translating song lyrics. An effective translation tool should have the flexibility to adapt and emphasize different poetic features in a translation search algorithm.

While the meaning of song lyrics is an important element of translation, it is not the only aspect of poetry. There has been significantly less work done involving the use of meter or rhyme in song translation, especially involving Modern Greek and Italian lyrics. For the aforementioned reason, this thesis aims to focus on creative computationally-driven translation techniques that differ from typical language models. It hopes to demonstrate why translating song lyrics should be approached as a problem distinct from translating “standard” poetry **[3]**. The Greek and Italian lyrics will be translated into English, but the approach would be similar for other target languages **[4]**.

Lyrics versus Poetry

The relationship between lyrics and poetry is tough to characterize. Although some would describe lyrics as a form of poetry, not all song lyrics are poetic and certainly not all poems fall into the category of lyrics. The main distinguishing feature of lyrics is that their form gets defined by the structure of the music they accompany. Unlike standard poetry, lyrics are primarily intended to be listened to and not be read. Despite the fact that some forms of poetry may be appealing to the ear when read aloud, poems are designed to be visually accessible. The fact that poems are presented in their “complete form” via a print medium also implies their intention to be read and reread at a pace comfortable for the individual audience members. Since song lyrics are often only presented in audio format, the listener does not have the luxury of following the words at a leisurely pace. The audience must discern and interpret a song’s lyrics in “real-time” as the music plays.

Of course, one can always go back and study the lyrics without the music, but this only happens when the listener deems that the lyrics are interesting enough to warrant further inspection. Some poems do get set to music after they have been around for many years, but this subset of poetry usually possesses “musical” qualities. Typically, lyricists, especially those who write for pop songs, craft their words such that they are “catchy” by taking advantage of the framework laid out by the music. The use of similar sounding words, especially at the end of musical phrases through rhyme, helps unify the lyrics and make them memorable. The lyrics also tend to avoid using overly-complicated figurative language because they are meant to be understood by the listener upon the first listen. While there are certainly a fair number of complex songs that require careful analysis to be fully appreciated, they still follow the form defined by the music. Additionally, it is much more common for lyrics to be overly formulaic, with a message that is explicitly expressed through simple and even cliché language.

Modern Greek and Italian Popular Music

Greece and Italy by nature of their proximity on the Mediterranean, have a rich shared culture and history. Modern Greek and Italian are both Indo-European languages with Ancient Greek and Latin influences. While the two languages are not mutually intelligible, their lexica both contain several words that originate from the same ancient roots. Greece and Italy also have compatible musical tastes as evidenced by the relatively frequent exchange of popular music across borders. Some of the most well-known Greek songs have become hits as Italian covers and vice versa. While it is not uncommon for an artist to have an international hit, a song has to be particularly compelling to cause an artist to translate and re-record a song in his/her native tongue. With the case of Greek and Italian musical exchanges, the translations serve as a way to convey a song's message to the public through meaningful music and lyrics. From anti-war protests to desperate pleas for love, the translations attempt to preserve the essence of their original versions **[5]**.

The main challenge of language transfer is that an entirely direct translation is impossible to achieve. In many cases there are words that simply do not exist in the repertory of another language, or if they do, may even have a conflicting meaning. Also, the word order and formation of a phrase follow different rules in other languages, such that the translation process involves much more than merely substituting one word from one language into another. Significant understanding about both the source and target languages' grammar and subtleties in vocabulary are therefore necessary for creating reliable translations. Even with expert knowledge of both languages, information is bound to get "lost in translation" when handling poetry. This is because some phrases simply become "unpoetic" when translated. Elements of rhyme or rhythm often need to be sacrificed in order to get the meaning correct. A common strategy for translators to create meaningful lyrics in the target language is to shift or create new details based the original to follow the poetic constraints. Some translations may be only loosely based on the source language text or even completely unrelated. A few of these techniques can be demonstrated with the following example:

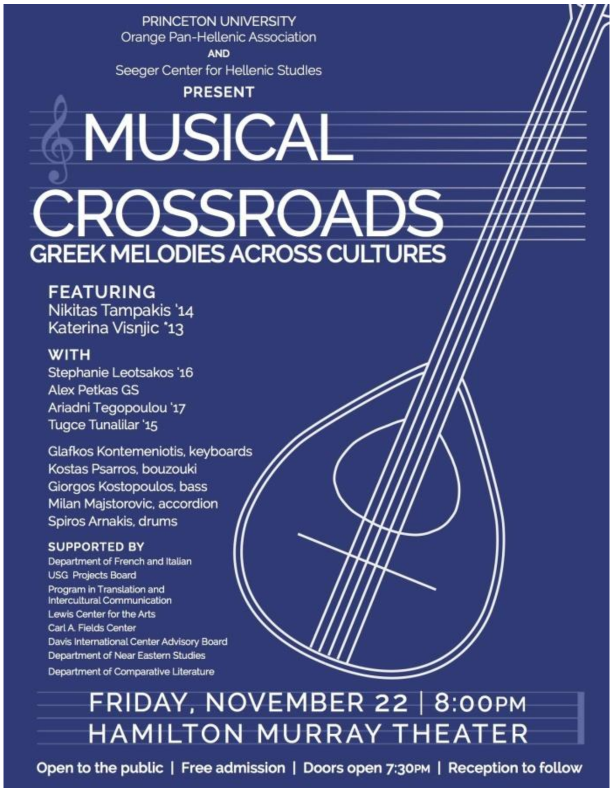
Frate Franc**e**sco partì una v**o**lta per oltrem**a**re
Fino alle t**e**rre di Babil**o**nia a predic**a**re

Στης Βαβυλ**ώ**νας τη γη με ρ**ά**σο αντί γι' ακ**ό**ντιο
Κινά ένας φ**ί**λος για ταξ**ί**δι υπερπ**ό**ντιο

Excerpt from "Il Sultano di Babilonia" written by Luisa Zappa and its corresponding Greek translation by Lina Nikolakopoulou

Presented above are the first two lines in the original Italian and Greek translation of the Angelo Branduardi song *Il Sultano di Babilonia e la prostituta* (The Sultan of Babylon and the Prostitute) **[6]**. The lyrics discuss the adventures of Saint Francis of Assisi on his journey to Babylon. Although both versions generally explain that Francis traveled overseas to Babylon to preach, they accomplish the task differently. In the Italian version "Friar Francis" is mentioned right in the opening of the verse (highlighted light blue), while in the Greek version, Francis is not mentioned in the first phrase and only is referred to as a "friend" in the second phrase. Babylon itself (highlighted green) is also contained in two separate lines. Yet, both versions fit within the same musical schema (stressed syllables bolded and underlined) as well as rhyme (highlighted red). The above example shows the liberty of word choice and order used by the translator in order to satisfy the poetic constraints.

Musical Crossroads



PRINCETON UNIVERSITY
Orange Pan-Hellenic Association
AND
Seeger Center for Hellenic Studies
PRESENT

MUSICAL CROSSROADS
GREEK MELODIES ACROSS CULTURES

FEATURING
Nikitas Tampakis '14
Katerina Visnjic '13

WITH
Stephanie Leotsakos '16
Alex Petkas GS
Ariadni Tegopoulou '17
Tugce Tunalilar '15

Glafkos Kontemeniotis, keyboards
Kostas Psarros, bouzouki
Giorgos Kostopoulos, bass
Milan Majstorovic, accordion
Spiros Arnakis, drums

SUPPORTED BY
Department of French and Italian
USG Projects Board
Program in Translation and Intercultural Communication
Lewis Center for the Arts
Carl A. Fields Center
Davis International Center Advisory Board
Department of Near Eastern Studies
Department of Comparative Literature

FRIDAY, NOVEMBER 22 | 8:00PM
HAMILTON MURRAY THEATER

Open to the public | Free admission | Doors open 7:30PM | Reception to follow

The poster for the Musical Crossroads Concert. Designed by Eric Shullman.

On Friday, November 22nd, 2013, I planned a concert program titled *Musical Crossroads: Greek Melodies Across Cultures*, which was presented in the Hamilton Murray Theater of Princeton University. The concert explored some of the most significant Greek cross-cultural musical exchanges not only involving Italy, but also France, India, Serbia, Turkey, and the US. A unique contribution of the concert was that songs were presented in multiple languages to emphasize the international aspect of the music presented. Members of the audience were surprised to hear that some of their favorite songs were actually cover versions and translated into their native tongue. The cultural movements explored through the concert provide a validation that language connects music to national identity and pride. Even if the melody of a song is imported from another country, effectively constructed translations of lyrics have the ability to "re-nationalize" a song into its target ethnicity.

On a personal note, I believe that the world's music should be accessible to everyone, and language barriers should not prevent it from being appreciated. While Chinese and English each have over one billion speakers, the majority of other languages have many fewer. Modern Greek, for example, has 15 million speakers worldwide and Italian has 60 million. While translations can help the user to “decrypt” the message of a song in an unfamiliar language, usually no effort is put in by the translator (human or machine!) to make the translation also singable. A song and its music can be appreciated in a more complete sense if it is also sung by the audience in a language they are comfortable using. As there are tens of thousands of songs written in Greek and Italian, an automatic translation method could be used preliminarily to guide the listener and allow them to sing along, since it is unlikely that a human has already translated it.

Automated and Computer-Assisted Translation

One would think that an ideal automatic lyrics translator would be able to take input lyrics from any song in text format and output the “best” translation possible in the target language. While a completely automated process would be useful in creating reasonable preliminary results with minimal effort on the user side, a more robust translator would be in the form of a toolkit that allows one to manipulate their results based on user-specified poetic and semantic features. The flexibility afforded by a toolkit also helps address the challenge of objectively defining an inherently subjective form of data. As both scenarios are of interest for music aficionados and professional translators alike, this thesis explores the process of generating a translation with and without human insight or intervention. Walking through the process of translating a few specific songs assisted by computational techniques provides useful insight for implementing a large-scale multi-lingual song translation application.

II. Data Collection

In addition to prior knowledge about popular Greek and Italian songs obtained through past performances and coursework at Princeton, the online database www.stixoi.info helped me compile many more Greek covers of Italian songs. The final dataset contains 28 popular songs with both Greek and Italian lyrics from the Second World War era through the present. Greek lyrics were downloaded from www.stixoi.info and from www.kithara.to, while the corresponding Italian lyrics were obtained from www.angolotesti.it and www.wikitesti.com. Italian songwriters contained in the set consist of Lucio Battisti, Angelo Branduardi, Paolo Conte, Lucio Dalla **[7]**, Jovanotti, and Nek, among others. Greek songwriters include Lavrentis Machairitsas, Thanos Mikroutsikos, Lina Nikolakopoulou, Mikis Theodorakis, and Dionysis Savvopoulos. As the data were compiled from multiple sources, I followed the collected lyrics and listened to each song carefully to screen out any inaccuracies and inconsistencies with the data. The complete list of songs is presented in Appendix A.

Despite having the same music and being credited as translations, the song lyrics vary significantly between their Greek and Italian editions. As previously mentioned, human translators and songwriters do not approach the language transfer problem of lyrics linearly and often "violate" the process that a translator would use with other media. Studying this "bilingual" music is useful because it provides insight into the poetic translation process, allowing one to analyze which elements of the text remain consistent across both languages and aids in conceptualizing a "language-independent" representation of song lyrics. Commonly observed deviations include:

- Shifting the words and concepts of the lyrics to different lines and verses from the original
- Maintaining only the topic or main theme of the original while generating new phrases to accompany the theme.
- Creating a completely different story that only happens to fit within the song structure of the original
- Using words that “sound” similar in different languages but are not necessarily semantically related.

Encoding Lyrics

The Greek and Italian lyrics are treated in terms of individual songs as well as a collection of data. By analyzing the entire dataset, we are able to analyze some overall data about the most frequently used words, the range of lexical choices, as well as overlap across both languages. As Modern Greek’s alphabet and Italian’s accented vowels are not part of the standard ASCII set, the UTF-8 encoding standard is used to represent the texts. Individual songs are presented in two formats:

- The original lyrics in .txt format - separate lines for each phrase and an empty line to separate distinct stanzas, such as verses and the chorus. The words can be matched against semantic networks to transfer the meaning of the words across languages.
- The lyrics divided by syllable in .csv format - each syllable is tagged with relevant pronunciation and performance information to analyze and compare the poetic similarities and differences in both language editions.

To explore potential differences between song lyrics and poetry, namely due to musical components, two tagging methods are used for creating the .csv files. One is based on examining meter and sonorities based on the textual format, as simply a poem. The other involves tagging the musical features of the lyrics. Prior research with computational poetry analysis and generation involves creating detecting the meter of a poem based on the natural pronunciation of the words contained in the poem. The hypothesis is that tagging the syllables with their corresponding musical stress yields more robust results than the results obtained without taking the musical context into account. This is rationalized by the fact that music causes syllable emphasis often to be distorted at the word level as well as the phrasal level. For example, inspect how the music alters the pronunciation of the following phrase from the Eagle's song *Hotel California*:

Correct spoken pronunciation: **Wel** - come to the Ho - **tel** Ca - li **-** **for** **-** nia
Pronunciation based on the music: Wel - **come** to the **Ho** - tel **Ca** - li **-** for **-** nia

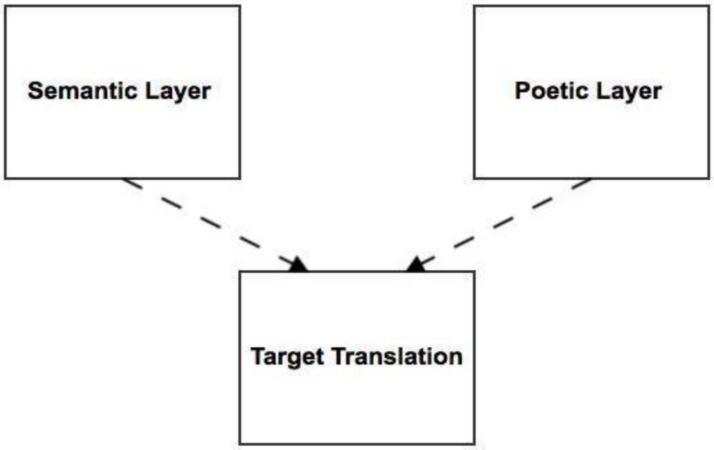
Lexical stress for multi-syllabic words is bolded and underlined

Without the music to guide the singer in performance, the presented line would normally be pronounced the first way, emphasizing the word "welcome" on the first syllable and the word "hotel" on the second syllable. However, performers of this song technically mispronounce these words when they accent "welcome" on the second syllable and "hotel" on the first syllable. The word "California" is stressed on two syllables, but the primary emphasis is on its first syllable Of course the change in prosody does not change the meaning of the lyrics, nor seems particularly unnatural in the context of the music.

Shifts in rhythmic stress also occur frequently when music is allowed to influence the shape of a phrase. This contrasts with standard poetry, which does not provide explicit phrasing instructions. In reciting poetry, typically an entire line is read aloud with each syllable receiving the same duration. Pauses are only taken when explicit punctuation marks indicate to do so, and the reader is given the liberty to choose to give a particular word emphasis. Musical phrasing in song lyrics allows for a much larger range of performance opportunities, such as the ability to place pauses within the middle of a word, encode pitch, and be flexible with the duration of each syllable. However, once the music is set, there is not as much room for altering this phrasing without "misinterpreting" the song.

Combining Poetics and Semantics

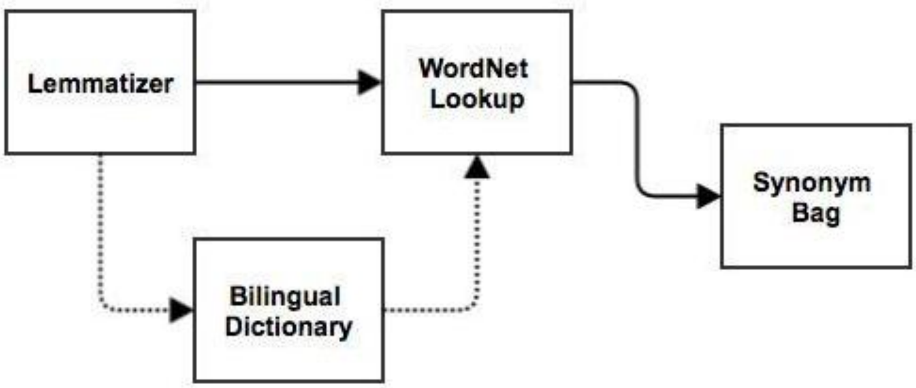
Treating the challenge of translating song lyrics as an information preservation problem, we seek to transfer as many of the original features as possible by leveraging the linguistic databases available for the languages of interest. Since the syntax of poetry is less-defined, we focus on extracting the poetic elements such as meter, music, and sounds, and semantic information represented mainly as concepts. The system aims to combine both of these elements in the filling in the syllable slots in the English translation. The following two sections will discuss the tools and process of constructing the semantic layer.



Song lyrics translation system design

III. Source Language Semantics

A diagram outlining the semantic layer of the translation system is shown below. It converts individual word tokens into their base form so that they can be searched and matched in two different lexical databases to map words and concepts into English. Semantically related words within the target language are then grouped together as possible candidates in the translation result.



Schematic for the semantic layer of the song lyrics translator system.

Available Resources

Many open-source tools primarily used for communicative-oriented natural language processing (NLP) tasks were used to analyze and reformat the test data. As Italian and especially Modern Greek are morphologically rich languages compared to English, an effective lemmatizer is essential for being able to extract the “root” form of the word, referred to as the lexeme. This is important because a word that appears in two different grammatical forms should be recognized as the same lexical unit despite not being identical. This process is relatively simple in English because each word unit can only exist in a very limited number of forms. For example, the adjective “beautiful” remains the same regardless of the noun it modifies. Yet in Italian, the equivalent adjective “bello” becomes “bella” when it is used to describe a feminine noun or “belli” when the noun is plural. Modern Greek adjectives uses different endings to encode gender and number, but also have different inflectional forms to encode the grammatical function it has in a given phrase. Recognizing these different representations and assigning a main entry, known as the lemma, is non-trivial and important for several translation-related tasks. It helps in accurately describing the diversity of the dataset through the quantity of unique tokens. Also, many bilingual dictionaries and semantic networks only recognize words in their lemma form.

The Institute for Language and Speech Processing of Greece hosts an online NLP tool that takes input text file written in Modern Greek, lemmatizes each word token, and tags the token with morphological information such as its part of speech, gender, number, tense, and inflection, where appropriate **[8]**. The output is an XML file hierarchically structured at the word level, line level, and the stanza level, which is easily parsable. The Italian Natural Language Processing Lab has an online tool called LinguA, which lemmatizes and annotates each word with its morphological and syntactic function for the Italian language and presents the results in the CoNLL format **[9]**. Although both tools yielded generally accurate matches, they are designed to work with processing standard prose. The song lyrics collected contain many instances of words being truncated, which is feature of colloquial language and is a technique employed to manipulate the rhythm of the phrase. Some the resulting lemmata were incorrect due to the truncation.

Based on the lemmatizer results across the entire dataset, the Greek data contain 5,916 tokens, 986 of which are unique. There are 6,645 Italian tokens with 1046 of them being unique.

In reality, the number of unique lemmata in each language should be less based on inaccuracies with the lemmatizer. First of all, both the Greek and Italian analyzer counts punctuation marks as word units, which are we are not interested in because line breaks separate phrases. The punctuation marks also have no semantic bearing and are not included in bilingual dictionaries. Additionally, the Italian lemmatizer treats each of the personal pronouns as separate entries, which does not occur in the Greek lemmatizer. Personal pronouns are arguably all the same lemma they all mark the subject of the sentence despite taking different forms to agree with number and perspective. In Italian, the consolidation also fails with some examples of truncation; *che* and the truncated *ch* are considered separate lemmata as well as some noun forms like *amore* and *amor*. While other forms of elision are actually required in standard written Italian, these examples are poetic manipulations. The Greek lemmatizer does even worse with the truncated and elided examples because standard writing never includes elision even though it is a feature of the spoken language. Some examples include *σε (se)*, *σ' (s')*, and, Σ' (S'), which all should representing the same form *σε*, a preposition meaning "in." Also, the word *είναι (einai),* meaning "it is" can be truncated at its head or tail, but the forms *είν' (ein)* and *'ναι (nai)* are misidentified as separate entries.

WordNet and bilingual dictionaries serve as useful tools for word transfer across languages. As extensions of the original English semantic network, other WordNet initiatives aim not only to connect related terms within a particular language, but also across languages. MultiWordNet is a project that mainly connects English and Italian concepts, but also contains support for Spanish, Romanian, Portuguese, Hebrew, and Latin **[10]**. BalkaNet is a project that connects the English and Greek WordNet, as well as Bulgarian, Serbian, Romanian, Turkish, and Czech **[11]**. The Greek WordNet has 18,000 synsets and the Italian WordNet has 67,000 synsets that correspond to the English WordNet **[12]**. If a lemma in Greek or Italian matches synset, then the semantic network of WordNet can produce a bag of similar words that are possible translation options. Also, the connection through the synset provides a network of Greek and Italian words.

Although both the Greek and Italian WordNets were matched to English synsets, one of the challenges that I faced involved using the same WordNet version. The current version of WordNet available for download is 3.0. However, BalkaNet is aligned with version 2.0 and MultiWordNet is aligned with version 1.6. While the WordNet group provides official mappings between versions, they only cover nouns and verbs and not adjectives or adverbs. Fortunately, the Natural Language Research Group in Spain provides mappings for all adjectives and adverbs as well. The group hosts mappings for across all pairs of WordNet versions from 1.5 to 3.0 **[13]**. When comparing Greek and Italian synsets, the Italian synsets are mapped from 1.6 to 2.0 in order to match the Greek synsets. 62,000 of the 67,000 original Italian synsets were successfully mapped. Whenever comparisons are made between Greek or Italian and English the original versions of the synsets get mapped to 3.0.

A drawback of the Greek WordNet, is that the semantic network data is relatively sparse compared to the English WordNet, which has over 110,000 synsets. This means that it is likely that frequently used words are missing from the dataset and will not be transferrable to the English WordNet. Another issue worth considering is polysemy, words matching multiple synsets. Polysemy is a relatively common phenomenon and is an issue because it is difficult to disambiguate the word sense of a particular word instance if it can represent several different concepts.

Greek and Italian Semantic Similarity

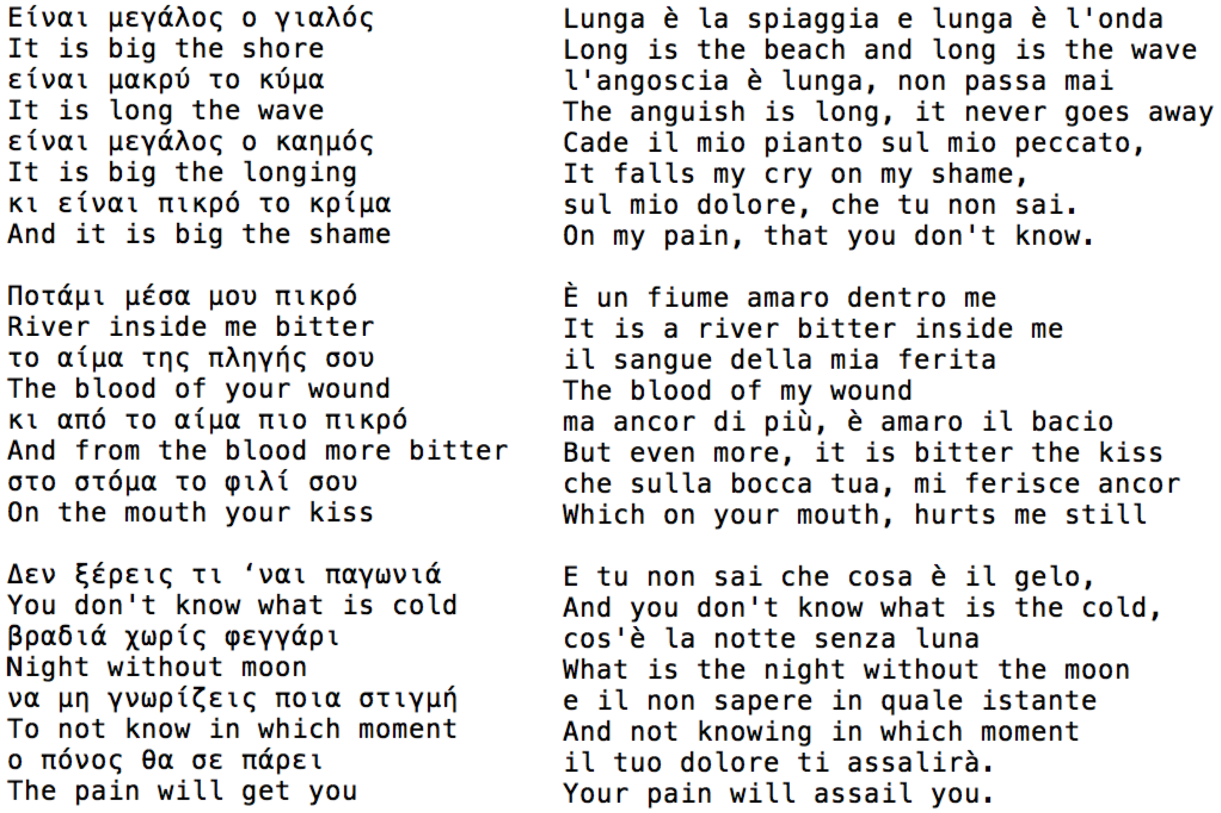
In the collected data, 454 of the 986 unique tokens, about 46% of the Greek data successfully matched with the Greek WordNet. Of the 454 matches, 249 were polysemous and led to a match of 1152 overall synset matches. In the Italian data, 802 of the 1046 tokens, 76% of the data matched with the Italian WordNet. 641 of the matches were polysemous with 2840 overall synset matches found. Of the overall synset matches in the Greek and Italian data 424 synsets were found in both sets.

There are multiple features of the data to explain the relatively low percentage of overlap (37% of the Greek synsets). Polysemy is the most obvious reason; multiple synsets per token approximately tripled the number of synsets in both datasets. The lack of overlap between many synsets polysemous words is an instance of "pruning" unrelated concepts, thus reducing the challenge of word sense disambiguation. Extraneous meanings are discarded if they do not exist in both languages' word to synset-pairs. Also, as previously mentioned the Greek WordNet is smaller than the Italian WordNet and simply does not have the same coverage. Another possibility is that the lack of synset overlap may just indicate that the songs do not hold much semantically in common with each other. We have seen that poetry is complicated to translate and it is fairly common for a songwriter to replace the ideas of the source to create new ones in the target language. These could be somewhat similar concepts or completely unrelated ones.

One other primary explanation for the lack of synset matches for related data is the occurrence of stop words. Stop words are frequently occurring words that are exist for more functional reasons than semantic. These are a language's prepositions, determinants, conjunctions, and pronouns. Although they are ubiquitous in English, Italian, and Modern Greek and possibly due to their large number of appearances, they are essentially meaningless. For this reason, words that are not nouns, verbs, adjectives, or adverbs are not included in the WordNets. What is tricky about the stop words in this context is that while most of them never have one of the 4 main part of speech tags in Modern Greek, many of them can occur in a "semantically-relevant" part of speech in Italian. Therefore a large number of stop words match with Italian synsets, but do not match with Greek synsets. For example, the Greek word *τι (ti)* and the Italian word *cosa* most frequently represent the determiner or pronoun *what* in English. In a different context, *cosa* also is a noun that means *thing*, which despite being vague, is a concept in WordNet. A list of previously compiled stop words for Greek and Italian were found online and were adjusted to omit words that are included in WordNet and can be found in Appendix B.

Specific Examples

Python scripts were written to parse the lemmatized data as well as perform the WordNet matches for the overall data set. The scripts can be applied to specific songs as well. To best illustrate the process we will study in depth the translation of a Mikis Theodorakis' song. In Greek the song is titled *Καημός (Kaimos - Longing),* and the Italian translation is titled *Un fiume amaro (A bitter river)*. The Greek lyrics are written by Dimitris Christodoulos and the Italian version was created by Sandro Tuminelli. Christodoulos is credited as co-author of the Italian version because the theme is the same in both languages. Below are the lyrics for both versions as well as a "direct" human translation:



Είναι μεγάλος ο γιαλός
It is big the shore
είναι μακρύ το κύμα
It is long the wave
είναι μεγάλος ο καημός
It is big the longing
κι είναι πικρό το κρίμα
And it is big the shame

Ποτάμι μέσα μου πικρό
River inside me bitter
το αίμα της πληγής σου
The blood of your wound
κι από το αίμα πιο πικρό
And from the blood more bitter
στο στόμα το φιλί σου
On the mouth your kiss

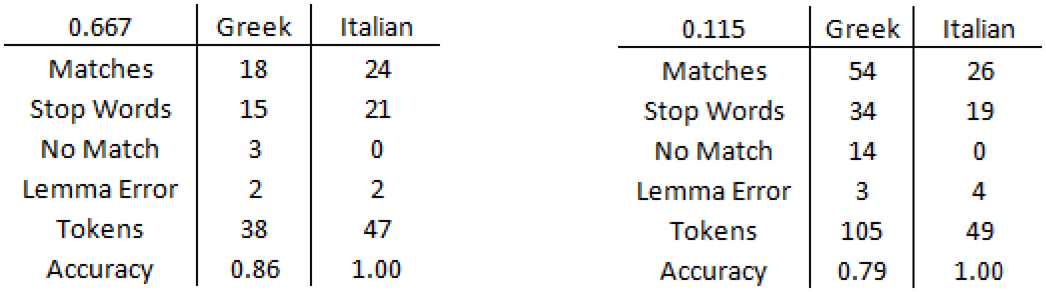
Δεν ξέρεις τι ‘ναι παγωνιά
You don't know what is cold
βραδιά χωρίς φεγγάρι
Night without moon
να μη γνωρίζεις ποια στιγμή
To not know in which moment
ο πόνος θα σε πάρει
The pain will get you

Lunga è la spiaggia e lunga è l'onda
Long is the beach and long is the wave
l'angoscia è lunga, non passa mai
The anguish is long, it never goes away
Cade il mio pianto sul mio peccato,
It falls my cry on my shame,
sul mio dolore, che tu non sai.
On my pain, that you don't know.

È un fiume amaro dentro me
It is a river bitter inside me
il sangue della mia ferita
The blood of my wound
ma ancor di più, è amaro il bacio
But even more, it is bitter the kiss
che sulla bocca tua, mi ferisce ancor
Which on your mouth, hurts me still

E tu non sai che cosa è il gelo,
And you don't know what is the cold,
cos'è la notte senza luna
What is the night without the moon
e il non sapere in quale istante
And not knowing in which moment
il tuo dolore ti assalirà.
Your pain will assail you.

Left: Kaimos - Greek original, Right: Un fiume amaro – Italian Translation
Rough English equivalent below each phrase.



0.667	Greek	Italian
Matches	18	24
Stop Words	15	21
No Match	3	0
Lemma Error	2	2
Tokens	38	47
Accuracy	0.86	1.00

0.115	Greek	Italian
Matches	54	26
Stop Words	34	19
No Match	14	0
Lemma Error	3	4
Tokens	105	49
Accuracy	0.79	1.00

Left: Kaimos/Un fiume amaro – WordNet matches and overlap for a semantically related song translation. Right: Giardini di marzo/Prin to telos – Results for a loosely-based Greek translation.

Upon inspection, it is apparent that the Italian translation contains more words per line than the original Greek does. This is confirmed by the larger number of tokens and synset matches in the data as seen in the WordNet chart above and to the left. The Greek data has 38 unique tokens, 18 of which match WordNet, while the Italian data has 47 tokens and 24 matches respectively. The WordNet match and overlap data included in the chart, quantify some the issues discussed in this section. The lemma error describes additional tokens that should not be included in the data because they are either punctuation or are not a unique lemma due to the lemmatizer failing to recognize elision. The data contain almost as many functional words as they do "significant" concept words. After removing the extraneous lemmata from the context, an adjusted accuracy score considers the percentage of matches in the remaining data. In both examples provided the Italian has 100% success rate in matching the important tokens to synsets. The Greek data is less successful and rather common words do not match. The title of the song *καημός (longing)* does not match and neither do *πικρός (bitter)* or *γιαλός (seashore).*

This is where the bilingual dictionary comes in. Using the PyGlossary tool[1], and the Babylon Greek-English dictionary[2], a searchable CSV file was compiled, mapping 50,000 Greek words into English. For the mismatches mentioned above, the unmatched words get mapped to English via the bilingual dictionary and then matched to the synsets in WordNet via lexical queries to the data. Overlap between the Italian and Greek versions increases to .684 because the word Greek word *πικρός* matches the synset accompanying the Italian word *amaro*.

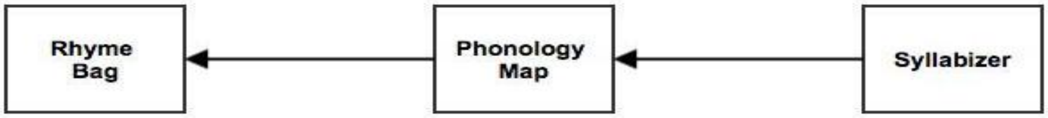
The overlapped synsets then each become a bag of synonyms which are the word senses, the synset's hyponyms, and co-hyponyms in the WordNet data. Costs are attached for appropriate semantic relatedness. For example, with the synset representing kiss (φιλί/bacio), a cost of 0 is attached to the word *kiss* since it is most directly correlated to the source language. A cost of 1 is given to the kiss's hyponym *smooch*, and a cost of 2 is given to co-hyponyms like *touch.*

[1] github.com/ilius/pyglossary

[2] www.babylon.com/free-dictionaries/languages/greek

IV. Source Language Poetics

The poetic layer of the translation system uses word pronunciation and musical data to extract the lyrics' syllable slots and meter as well as capture the sounds of the original. When searching for a translation, we need to know which words and syllables to emphasize and would like to preserve some of the sonority features of the original. Words get split into syllables, get tagged (manually) with musical data, and mapped across the phonologies or sound systems of two languages. To be precise about phonotactics and syllables, words in song lyrics are defined by their syllables, which receive a pitch when voiced or sung. At the center of every syllable is its vowel nucleus, which can be surrounded on either end by a single or compound consonant head and tail **[14]**. There is significant overlap in the phonemic inventories of languages, particularly with "simple" vowels and consonants, meaning that the sounds in many languages can be produced by humans even without understanding their meaning.



Overview of the Poetic Layer

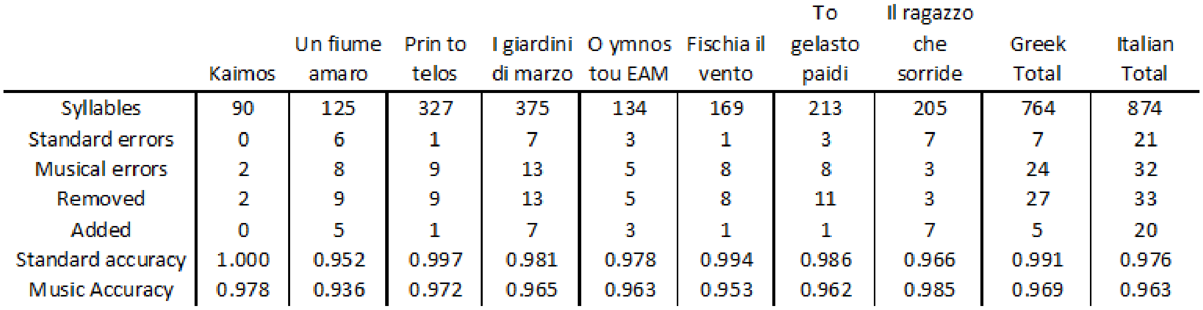
Tagging Process

A Greek[3] and Italian syllabizer[4] does a decent preliminary job of splitting the words into their appropriate syllables, which is not a trivial process particularly when two vowels appear adjacent to each other **[15]**. One drawback of these systems, which are designed primarily for

[3] nlp.ilsp.gr/soaplab2-axis/
[4] www.sillabare.it

hyphenating words that split across two lines of text, is that the words are treated independently of their context. Each word boundary is treated as a syllable break, which makes sense in text form, but it does not represent the correct pronunciation, often even in its standard spoken context since words that end and start with similar vowels often elide. In music, songwriters are flexible in deciding whether or not to combine the two syllables at a word boundary into one based on their artistic preference. Because music can combine or separate any two syllables, syllabification of song lyrics is not predictable from the text alone. Sometimes the "natural" elision is not practiced so that a desired syllable pattern is met.

To evaluate the accuracy of the syllabizers and observe how problematic elision is, we compare the errors found in regular speech syllabification with the ones that the music defines. The syllable-separated lyrics were first read aloud normally "as a poem" to see if the syllable divisions made sense. They were then checked closely by following along with the music to see if word boundaries should be collapsed into a single syllable. Each correction by addition or removal of a syllable is marked as an error.



	Kaimos	Un fiume amaro	Prin to telos	I giardini di marzo	O ymnos tou EAM	Fischia il vento	To gelasto paidi	Il ragazzo che sorride	Greek Total	Italian Total
Syllables	90	125	327	375	134	169	213	205	764	874
Standard errors	0	6	1	7	3	1	3	7	7	21
Musical errors	2	8	9	13	5	8	8	3	24	32
Removed	2	9	9	13	5	8	11	3	27	33
Added	0	5	1	7	3	1	1	7	5	20
Standard accuracy	1.000	0.952	0.997	0.981	0.978	0.994	0.986	0.966	0.991	0.976
Music Accuracy	0.978	0.936	0.972	0.965	0.963	0.953	0.962	0.985	0.969	0.963

Syllabizer results for 4 "pairs" of Greek and Italian Songs

The main “standard” errors with the syllabizer involved failing to separate adjacent vowels into separate words when necessary. This occurred more often with the Italian tool than the Greek. In Italian, frequently-occurring words like *mio (mine), ogni (every), spiaggia (beach)* would need one extra syllable spot. The only “standard” errors with the Greek syllabizer involved acronyms like ΕΑΜ or ΕΛΑΣ, which are typically pronounced as written with two syllables and not one. Some vowels get completely discarded, while others become diphthongs. Despite these errors, without any human input, the syllabizers are highly accurate, which suggests that the syllabification process does not necessarily need to checked manually.

The addition of musical emphases by syllable was also done manually, and will likely still need human tagged meta-information in future work. Musical information retrieval is a challenging area of research and does not have much success in obtaining precise meter information from an audio file. I also wanted to make sure that this information was accurate to demonstrate clearly that the music is more important than the words alone in translation song lyrics. With the corrected syllable data, each syllable is formatted into a .csv file tagged with lexical and musical elements. The features include whether it is the lexical or phrasal musical stress, if it is the highest or lowest musical note, and the syllable’s note duration. Lexical emphases are dropped if they are truncated for a non lexical emphasis sound. The file is structured at the syllable-, phrasal-, and sectional- level to allow for comparisons within and across a verse or chorus. It also allows for analysis between the two language versions.

With the music meta-data, we can determine the meter of the phrase by calculating a score taking all the tagged musical emphases into account. The scoring calculation is explained on the next page.

Music stress:

- S – long note on strong beat + 2
- s – short note on strong beat + 1.5
- l – long note not on a strong beat +1
- w – short note on weak beat +.5
- n – no rhythmic emphasis 0

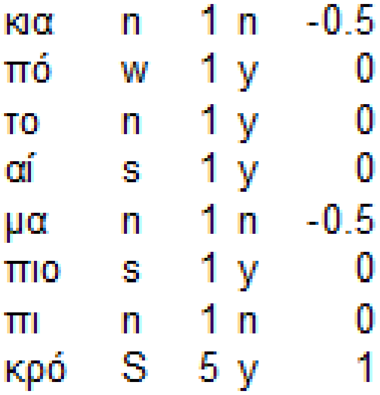
Number of notes per syllable:

Most syllables are sung on one note, so any additional pitches attached to a syllable bring attention to it +.5 for each additional pitch

Highest or lowest note 1 * magnitude

Pitch maxima and minima:

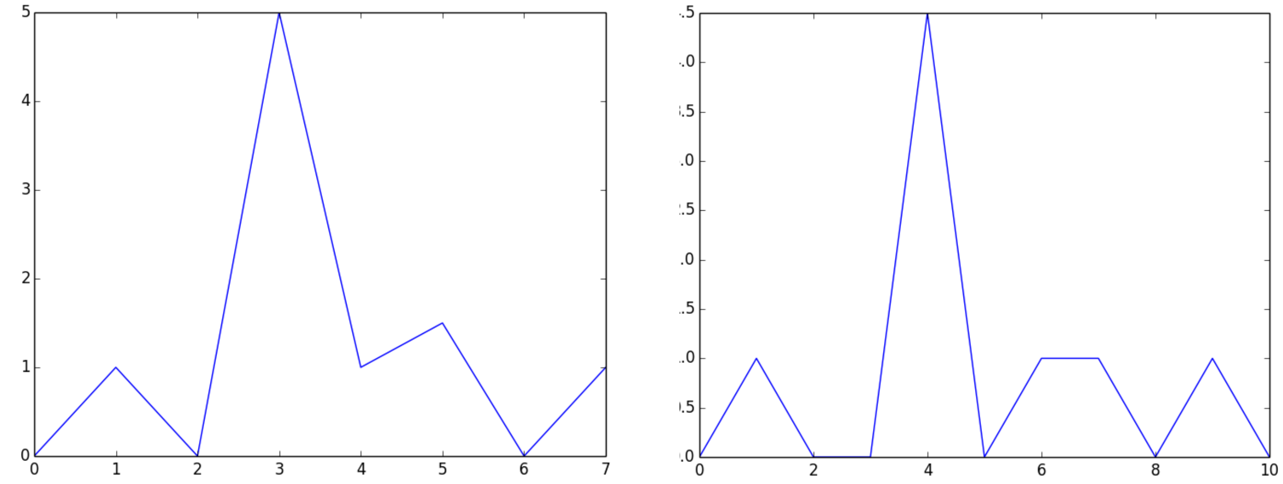
The highest and lowest notes in a phrase stand out, especially if there is only one instance of them. +1/number of notes sharing the highest or lowest pitch



κια	n	1	n	-0.5
πό	w	1	y	0
το	n	1	y	0
αί	s	1	y	0
μα	n	1	n	-0.5
πιο	s	1	y	0
πι	n	1	n	0
κρό	S	5	y	1

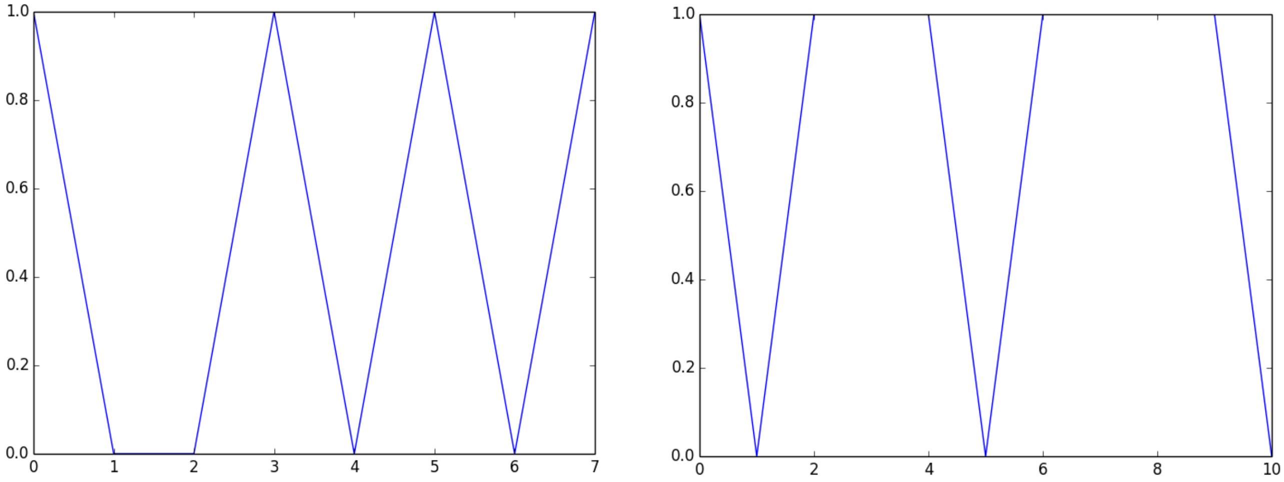
Sample of the tagged data. Columns as follows:
1st: syllable, 2nd: rhythmic emphasis, 3rd: number of pitches,
4th: is lexical emphasis?, 5th: Pitch maximum and minimum

Lyrics and Poetry Meter Visualization



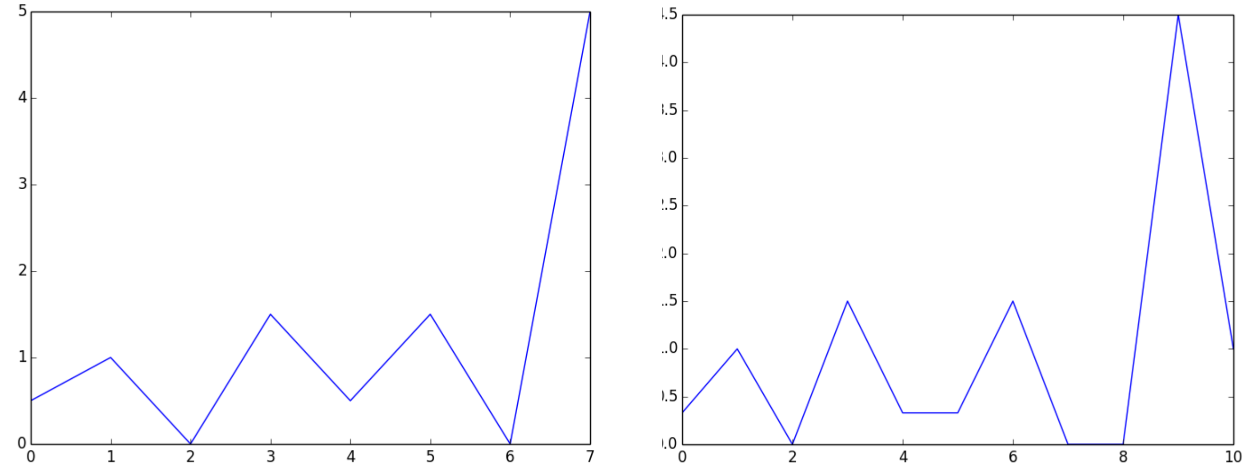
Ei - **nai** me - **ga** - los **o** gia - **los** Lun - **ga** è la **spiag**-gia **e** **lun** –g’è **l’on** - da

Musical emphasis score for a phrase during a stanza of Kaimos/Fiume amaro. Left: Greek original, Right: Italian Translation



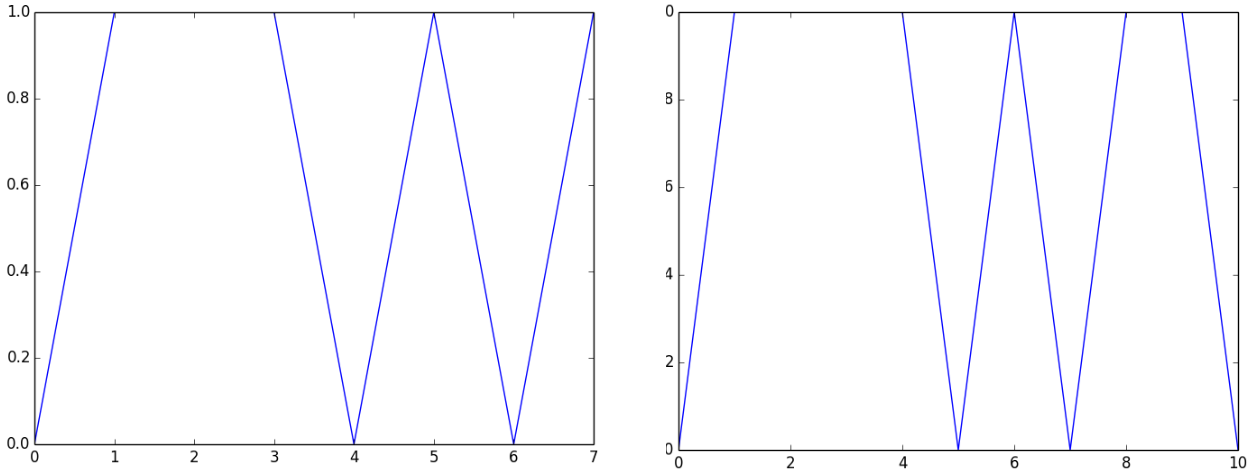
Ei - nai me - **ga** - los **o** gia - **los** **Lun** - ga **è** **la** **spiag**-gia **e** **lun**-**g’è** **l’on** - da

Lexical emphasis score (without music) for a phrase during a stanza of Kaimos/Fiume amaro. Left: Greek original, Right: Italian Translation



ki’a - **po** to **ai** - ma **pio** pik - **ro** m’an - **cor** di **più** è a - **ma** ro il **ba**-cio

Musical emphasis score for a phrase during a stanza of Kaimos/Fiume amaro. Left: Greek original, Right: Italian Translation



ki’a - **po** **to** **ai** - ma **pio** pik - **ro** m’an - **cor** **di** **più** **è** a - **ma** ro **il** **ba**-cio

Lexical emphasis score (without music) for a phrase during the chorus of Kaimos/Fiume amaro. Left: Greek original, Right: Italian Translation

The scoring procedure makes it clear which syllable or word the songwriter wants emphasized the most. The multiple musical features show a higher ability to emphasize syllables at different levels. Since poetry results only take lexical emphasis into account, the meter is not as clearly defined as it is accompanied by the music. While the music shows a consistent weak-strong pattern for 4 beats in a phrase, the poetic information alone is insufficient in determining the appropriate meter.

Mapping Sounds Across Languages

Besides rhythm, song lyrics are notable for the way the sounds interact with each other. Beyond music, Italian is often considered a "musical" language simply based on the sounds that are part of its phonemic inventory **[15]**. Although phonotactics, or the way syllables can be constructed and combined to form words, differs between languages, all sounds in English can be produced or at least approximated by Greek and Italian speakers. Even though the sound units of human language are well documented, they are mainly studied as pieces of a mono-linguistic system. Very little research exists on comparing phonetic systems across languages, particularly as a tool in song translation. This has the potential to be a useful tool because a translation system can select words not only represent the same concept, but also "sound nice." Also phone-based search can help identify words that are etymologically similar.

In order to translate "sounds" across languages we need to be able to have representations of the sound data. The alphabets of Modern Greek and Italian are sufficient for this task because they graphemes correspond mechanically to phonemes, the sound units of a language **[16]**. This is not the case for English however, where there is a many-to-many mapping of letters to sound units ('gh' is a different phoneme in rough and through and has to be "learned" as an exception). To serve the purpose mapping words to sounds in English, Carnegie Mellon has compiled 125,000-word English dictionary containing the standard spelling of a word and its corresponding phonetic transcription (for a North-American speaker)[5]. The data are organized as follows:

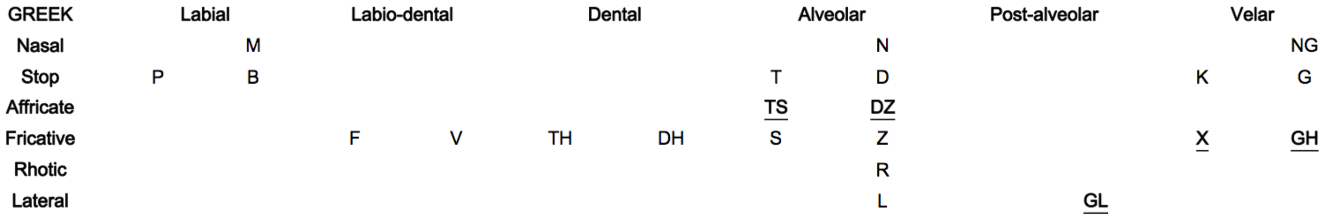
COOPERATE K OW0 AA1 P ER0 EY2 T

The database represents individual sound units via the ARPAbet, which was first developed during the 1970s at the dawn of significant computer voice synthesis research **[17]**. The ARPAbet is an encoding system for the 39 main phonemes of the English language containing 24 consonants and 15 vowels. The vowels each represent the main body of a syllable and have stress information '0', '1', '2' appended. The '0' represents that the vowel is unstressed, the '1' is for the main lexical stress, and the '2' is for a secondary lexical stress. The format allows machines to easily parse and process the pronunciation data in speech synthesis systems. The complete ARPAbet is included in Appendix C.

The sounds that form the phonology of a language are a form of poetry that play a large role in the songwriting/translating process. The purpose of this mapping is to analyze similarities in the sounds employed in both song editions as well as match them with English sounds. The challenge with mapping phonologies to English is that some songs in Greek/Italian sounds do not exist in English while some common sounds in English do not exist in Greek/Italian. To make sure there is a possible match for every word in the CMU speaking dictionary, a scoring system is described for favoring matches that most directly correspond, but also allowing for approximate matches to theoretically be found for all English words.

[5] www.speech.cs.cmu.edu/cgi-bin/cmudict

Modern Greek Consonant Phonology



GREEK	Labial		Labio-dental		Dental		Alveolar		Post-alveolar		Velar	
Nasal		M						N				NG
Stop	P	B					T	D			K	G
Affricate							TS	DZ				
Fricative			F	V	TH	DH	S	Z			X	GH
Rhotic								R				
Lateral								L		GL		

The Modern Greek representation of its consonants in ARPAbet format is displayed above. 17 of the 24 basic English phonemes for consonants are part of the Greek system **[18]**. There are 5 consonant sounds in Greek which does not exist in English and are underlined in the chart. New symbols are created to fully encode the Greek information before searching for approximate matches. The cost function is defined as follows for English approximants of Greek is as follows:

TS ->T S (2 units) +0, TS -> CH +1 GH -> G +1, GH -> Y +1

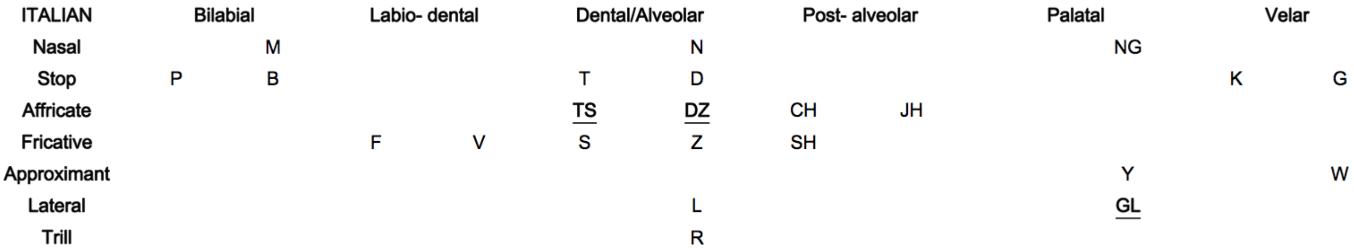
GL -> L Y +0, GL -> L IY +.5, GL -> Y +1, GL -> L + 1.5

DZ ->D Z (2 units) +0, DZ -> JH +1 X -> K +1, X -> HH +1

The English HH, Y, CH, JH sounds are covered with this mapping. For full coverage of the remaining three English consonant phonemes, the costs for Greek approximants of English are:

Z -> ZH +1 UW (vowels discussed later) -> W +1 S -> SH +1

Italian Consonant Phonology

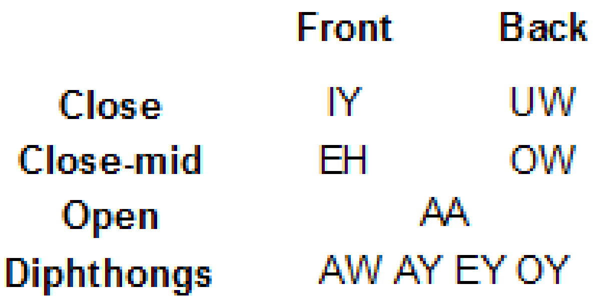


ITALIAN	Bilabial		Labio- dental		Dental/Alveolar		Post- alveolar		Palatal		Velar	
Nasal		M				N				NG		
Stop	P	B			T	D					K	G
Affricate					TS	DZ	CH	JH				
Fricative			F	V	S	Z	SH					
Approximant										Y		W
Lateral						L				GL		
Trill						R						

20 of Italian's 23 consonants match directly with 20 of English's 24 consonants. The three consonant sounds that are unique to Italian also exist and Greek and use the same mapping cost to English. The 4 English sounds that Italian does not have are DH, HH, TH, and ZH. The costs are as follows:

D -> DH +1 T -> TH +1 HH -> +1 (blank) Z -> ZH +1

Greek and Italian Vowel Phonology



	Front	Back
Close	IY	UW
Close-mid	EH	OW
Open	AA	
Diphthongs	AW AY EY OY	

While some would argue that the Italian phonology has the open-mid back vowel (AO in English ARPAbet), this is not true for all Italian speakers, since it is very similar to the OW sound **[19]**. Besides that distinction, the Greek and Italian vowel inventories are the same. This means that both languages match 9 of the 14 English vowel sounds. The rest need to be approximated as follows:

AA -> AE +1, AA -> AH +1 EH -> ER +1 EH R -> ER +0

IY -> IH +1 UW -> UH +1 OW -> AA +1, OW -> AO +1

V. Evaluation and Future Work

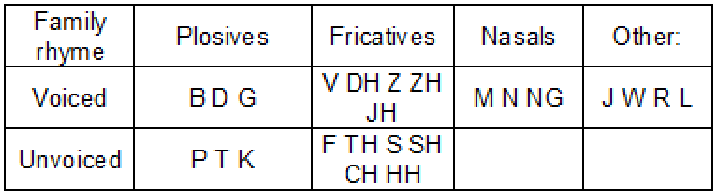
The bulk of this paper discusses techniques one can use to extract semantic and poetic information from song lyrics. Although the goal was to be able to explore and evaluate results of entire song lyrics, the data extraction phase proved to be more challenging than anticipated. What this paper does accomplish is that it demonstrates that song lyrics should be treated as a class distinct from standard poetry as the musical information provided very different meter and phrasal emphases results. It also describes a way to utilize linked WordNet for translation of concepts and a way to cope for less-developed networks via a bilingual dictionary. The phonology mapping is certainly novel and is an important cornerstone of a translation system based on sonorities.

Since I devised a scoring system for semantic relatedness and phonology similarity, the main challenge lies in combining the meaning and music. A path one could explore would involve evolutionary algorithms or greedy edit distance algorithms based on these scores in a manner similar to what is described in Manurung's dissertation on poetry generation **[20]**. Homophonic translation is also a promising direction. Overall, I'm happy that this thesis gave me the opportunity to combine my computational and creative sides. I love writing songs and I explored how linguistic networks can help me write even better songs. The remainder of the section describes other possible future work.

Rhyming Strategies

A basic rhyming dictionary was implemented built from the CMU pronunciation data that can perform perfect rhyme searches. It can also be fed related strings to complete the other rhyme searches. It can be used returns a list of words that rhyme with any given word ranked in the following order based on practices used by professional songwriters **[21]**:

- Perfect rhyme – Two words have a perfect rhyme when the lexical stress matches and the tail of the world after the accent is identical. Example: tricycle and icicle
- Family rhyme – Almost identical to the perfect rhyme except the consonants that follow the stressed syllable are in the same consonant family. Example: fairy and daily



Family rhyme	Plosives	Fricatives	Nasals	Other:
Voiced	B D G	V DH Z ZH JH	M N NG	J W R L
Unvoiced	P T K	F TH S SH CH HH		

Each square displays English phonemes that are in the same family and almost rhyme perfectly.

- Additive rhyme – A perfect rhyme except the last syllable has an extra consonant tail. Example: Fry and my are perfect rhymes, fry and might are additive rhymes
- Subtractive rhyme – The reverse of additive rhyme. The last syllable has part of its tail missing. Example: Might and fright are perfect, might and fry are not.
- Assonance – When all of the vowel sounds after the lexical stress are identical, but not the consonants. Example: matter and amber
- Consonance – When consonants are identical but vowels are not. Example: spike and spook

Homophonic Translation

Soramimi is a Japanese word meaning “misheard” and is a term applied to refer to the act of intentionally mishearing lyrics, particularly of those written in a different language. The use of technique is an interesting approach to lyrics translation because it involves mapping the sounds of the original song as directly as possible to words in the target language. This technique has been applied with a usually humorous result, but has never been approached as a computer matching problem. It involves homophonic translation, which is taking the sounds of one language and turning them into words. Some songs in the dataset use this technique in a way. “Come Monna Lisa” in became “Μην ορκίζεσαι” in Greek because of phonological, not semantic similiarities.

Stop Words and Bigrams

Stop words are the most frequently occurring words in language and they are useful in filling out the syllable slots in song lyrics as well. However, they are not included in the WordNets, and do not get transferred across languages. As a result statistical language models can help fill in the gaps surrounding the “important” words that do get mapped into the target language. A bigram tool showing which words are most likely to occur directly before and after a particular word would help populate the syllable of a line as well as improve the overall cohesiveness of a phrase. One way to do this could be based on collocation scores with appropriate corpuses **[22, 23, 24]**.

Works Cited

[1] Elytēs, Odysseas, Edmund Keeley, and Geōrgios P. Savvidēs. *The Axion Esti*. Pittsburgh: U of Pittsburgh, 1974. 67.

[2] Menin, Roberto. *Tradurre la canzone d'autore*. Eds. Giuliana Garzone, and Leo Schena. CLUEB, 2000.

[3] Genzel, Dmitriy, Jakob Uszkoreit, and Franz Och. "Poetic statistical machine translation: rhyme and meter." Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010.

[4] Ramakrishnan A, Ananth, Sankar Kuppan, and Sobha Lalitha Devi. "Automatic generation of Tamil lyrics for melodies." Proceedings of the Workshop on Computational Approaches to Linguistic Creativity. 2009.

[5] Valenti, Gianluca. "Da canzoni (e poesie) a canzoni: adattamenti metrici nelle traduzioni di Fabrizio de André." *Signa: Revista de la Asociación Española de Semiótica* 22 (2013).

[6] Schweighofer, Ingrid. *Il contributo della canzone italiana alla glottodidattica attuale – Angelo Branduardi*. Diss. uniwien, 2010.

[7] Macale, Maurizio, and Lucio Dalla. *Caro amico ti scrivo. Da Il cielo a Ciao", Foggia, Bastogi, 2000*. ISBN 88-8185-281-0.

[8]Prokopidis, Prokopis, Byron Georgantopoulos, and Haris Papageorgiou. "A suite of NLP tools for Greek." *Proceedings of the 10th International Conference of Greek Linguistics (ICGL 2011), Komotini, Greece*. 2011.

[9] Montemagni S. (2013), "*Tecnologie linguistico-computazionali e monitoraggio della lingua italiana*". In Studi Italiani di Linguistica Teorica e Applicata (SILTA) Anno XLII, Numero 1, pp. 145-172.

[10] Tufis, Dan, Dan Cristea, and Sofia Stamou. "BalkaNet: Aims, methods, results and perspectives. a general overview." *Romanian Journal of Information Science and Technology* 7.1-2 (2004): 9-43.

[11] Magnini, Bernardo, and Carlo Strapparava. "Costruzione di una base di conoscenza lessicale per l'italiano basata su WordNet." *Proceedings of the 28th International Congress of the Societ a Linguistica Italiana, Palermo, Italy*. 1994.

[12] Fellbaum, Christiane. WordNet. Blackwell Publishing Ltd, 1999.

[13] Daudé, Jordi, Lluis Padro, and German Rigau. "Mapping wordnets using structural information." Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2000.

[14] Celata, Chiara, and Basilio Calderone. "Restrizioni fonotattiche, pattern lessicali e recupero delle regolarità morfologiche. Convergenze computazionali e comportamentali."

[15] Maturi, Pietro. *I suoni delle lingue, i suoni dell'italiano: introduzione alla fonetica*. Il mulino, 2006.

[16] Chalamandaris, A. Raptis, S., & Tsiakoulis, P. (2005). Rule-based grapheme-to-phoneme method for the Greek. In *INTERSPEECH-2005* pp. 2937–2940.

[17] Klatt, Dennis H. "Review of the ARPA speech understanding project." The Journal of the Acoustical Society of America 62 (1977): 1345.

[18] Arvaniti, Amalia. "Greek Phonetics." Journal of Greek Linguistics 8 (2007): 97-208.

[19] Canepari, Luciano. *Dizionario di pronuncia italiana*. Zanichelli Editore, 1999.

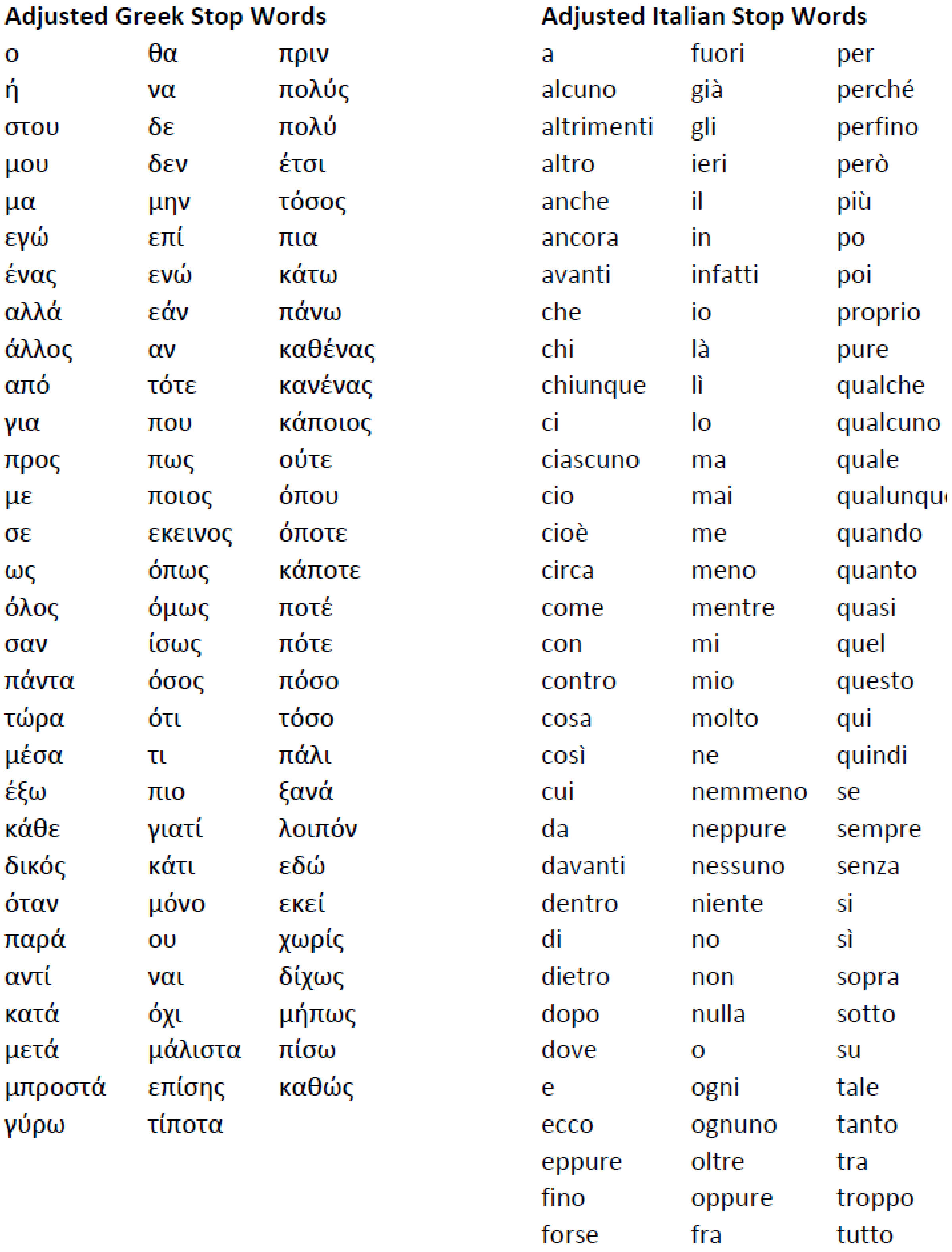
[20] Manurung, Hisar. "An evolutionary algorithm approach to poetry generation." (2004).
[21] Pattison, Pat. *Writing better lyrics*. Writer's Digest Books, 2009.
[22] Cresti, Emanuela. *Corpus di italiano parlato: Introduzione*. Vol. 1. Accademia della Crusca, 2000.
[23] Davies, Mark. (2014) The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.
[24] Strapparava, Carlo, Rada Mihalcea, and Alberto Battocchi. "A Parallel Corpus of Music and Lyrics Annotated with Emotions." *LREC*. 2012

Appendix A: List of Songs

Greek Song Title	Italian Song Title	Composer
Ο χρόνος που μετράει	L'anno che verrà	Lucio Dalla
Ο ύμνος το ΕΑΜ	Fischia il vento	Matvei Blanter
Συννεφιασμένε μου ουρανέ	La canzone del sole	Lucio Battisti
Άρνηση	Sogno di libertà	Mikis Theodorakis
Τέρμα η ιστορία	Lascia che io sia	Nek
Εγώ για σένα	Ancora vivo	Gianni Bella
Το γέλαστο παιδί	Il ragazzo che sorride	Mikis Theodorakis
Μάρκος και Άννα	Anna e Marco	Lucio Dalla
Στη γιορτή της αυγής	Alla fiera dell'est	Angelo Branduardi
Μην ορκίζεσαι	Come Monna Lisa	Mango
Φίλα με ακόμα	Baciami ancora	Jovanotti
Ο καημός	Un fiume amaro	Mikis Theodorakis
Πριν το τέλος	I giardini di Marzo	Lucio Battisti
Ο Σουλτάνος της Βαβυλώνας	Il Sultano di Babilonia	A. Branduardi/A. Parente
Τα λέμε	Bastardo	Gigi D'Alessio
Σκέψου καλά	Laura non c'è	Nek
Χρόνια	Anni	Paolo Conte
Λάμπει	Dal loggione	Paolo Conte
Τερατάκια τσέπης	Vanità di vanità	Angelo Branduardi

Σαν Αμερικάνος	Tu vuo’ fa’ l’americano	Renato Carosone
Τι να πω	E penso a te	Lucio Battisti
Αντίο αντίο αγάπη	Amara terra mia	Domenico Modugno
Ε και	La differenza tra me e te	Tiziano Ferro
Μια μέρα δάνεικη	Quanti anni hai	Vasco Rossi
Μάυρο μαργαριτάρι	Dolcenera	Fabrizio de Andre
Η αγάπη της ζωής μου	Cose della vita	Eros Ramazzotti
Άννα	Anna	Lucio Battisti
Μια πίστα από φώσφορο	Il canto di un’Eneide diversa	Thanos Mikroutsikos

Appendix B: Greek and Italian Stop Words



Adjusted Greek Stop Words

ο
ή
στου
μου
μα
εγώ
ένας
αλλά
άλλος
από
για
προς
με
σε
ως
όλος
σαν
πάντα
τώρα
μέσα
έξω
κάθε
δικός
όταν
παρά
αντί
κατά
μετά
μπροστά
γύρω

θα
να
δε
δεν
μην
επί
ενώ
εάν
αν
τότε
που
πως
ποιος
εκεινος
όπως
όμως
ίσως
όσος
ότι
τι
πιο
γιατί
κάτι
μόνο
ου
ναι
όχι
μάλιστα
επίσης
τίποτα

πριν
πολύς
πολύ
έτσι
τόσος
πια
κάτω
πάνω
καθένας
κανένας
κάποιος
ούτε
όπου
όποτε
κάποτε
ποτέ
πότε
πόσο
τόσο
πάλι
ξανά
λοιπόν
εδώ
εκεί
χωρίς
δίχως
μήπως
πίσω
καθώς

Adjusted Italian Stop Words

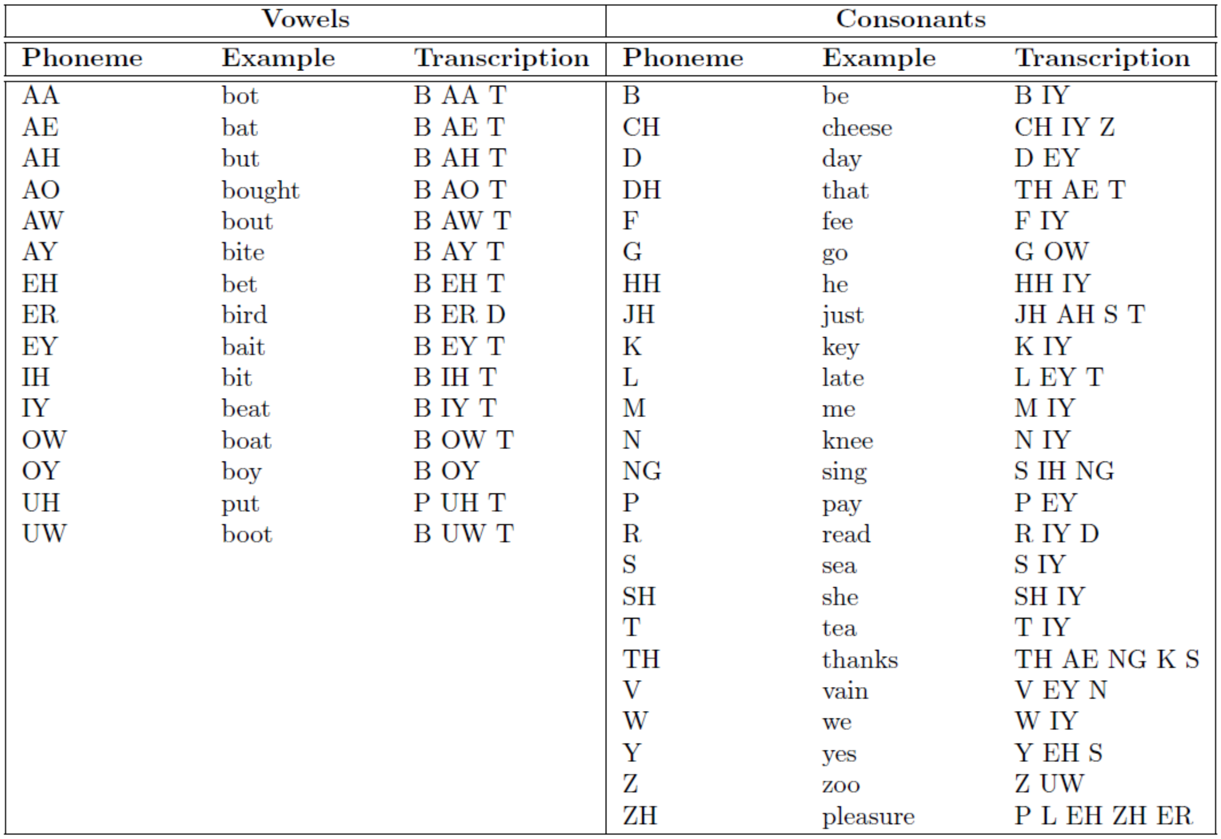
a
alcuno
altrimenti
altro
anche
ancora
avanti
che
chi
chiunque
ci
ciascuno
cio
cioè
circa
come
con
contro
cosa
così
cui
da
davanti
dentro
di
dietro
dopo
dove
e
ecco
eppure
fino
forse

fuori
già
gli
ieri
il
in
infatti
io
là
lì
lo
ma
mai
me
meno
mentre
mi
mio
molto
ne
nemmeno
neppure
nessuno
niente
no
non
nulla
o
ogni
ognuno
oltre
oppure
fra

per
perché
perfino
però
più
po
poi
proprio
pure
qualche
qualcuno
quale
qualunqu
quando
quanto
quasi
quel
questo
qui
quindi
se
sempre
senza
si
sì
sopra
sotto
su
tale
tanto
tra
troppo
tutto

Adapted from 29-language stop word data: https://code.google.com/p/stop-words/

Appendix C: ARPAbet



Vowels			Consonants		
Phoneme	Example	Transcription	Phoneme	Example	Transcription
AA	bot	B AA T	B	be	B IY
AE	bat	B AE T	CH	cheese	CH IY Z
AH	but	B AH T	D	day	D EY
AO	bought	B AO T	DH	that	TH AE T
AW	bout	B AW T	F	fee	F IY
AY	bite	B AY T	G	go	G OW
EH	bet	B EH T	HH	he	HH IY
ER	bird	B ER D	JH	just	JH AH S T
EY	bait	B EY T	K	key	K IY
IH	bit	B IH T	L	late	L EY T
IY	beat	B IY T	M	me	M IY
OW	boat	B OW T	N	knee	N IY
OY	boy	B OY	NG	sing	S IH NG
UH	put	P UH T	P	pay	P EY
UW	boot	B UW T	R	read	R IY D
			S	sea	S IY
			SH	she	SH IY
			T	tea	T IY
			TH	thanks	TH AE NG K S
			V	vain	V EY N
			W	we	W IY
			Y	yes	Y EH S
			Z	zoo	Z UW
			ZH	pleasure	P L EH ZH ER

Downloaded from UPenn Department of Linguistics Website
http://fave.ling.upenn.edu/downloads/ARPAbet.pdf