Learning Phonology by Machines

Nikitas N. Karanikolas
Dept. of Informatics and Computer Engineering
University of West Attica, Athens, Greece
nnk@teiath.gr
http://users.teiath.gr/nnk/

What is phonology [1]

- **Phonology** is a branch of linguistics concerned with the systematic organization of sounds in languages. It has traditionally focused largely on the study of the systems of phonemes in particular languages (and therefore used to be also called *phonemics*, or *phonematics*), but it may also cover any linguistic analysis either at a level beneath the word (including syllable, onset and rime, articulatory gestures, articulatory features, mora, etc.) or at all levels of language where sound is considered to be structured for conveying linguistic meaning.

Phonology vs Phonetics [1]

- Phonology is often distinguished from *phonetics*. While phonetics concerns the physical production, acoustic transmission and perception of the sounds of speech, phonology describes the way sounds function within a given language or across languages to encode meaning. For many linguists, phonetics belongs to descriptive linguistics, and phonology to theoretical linguistics.

Phoneme [2]

- A **phoneme** (/ˈfoʊniːm/) is one of the units of sound (or gesture in the case of sign languages, see chereme) that distinguish one word from another in a particular language. For example, in most dialects of English, the sound patterns /θʌm/ (*thumb*) and /dʌm/ (*dumb*) are two separate words distinguished by the substitution of one phoneme, /θ/, for another phoneme, /d/. In many other languages these would be interpreted as exactly the same set of phonemes (i.e. /θ/ and /d/ would be considered the same).

Some phonemes does not exist in every language

- GR cap – GR low – ALB – IPA

- Δ – δ – Dh – ð
- Dhjaku
- Διάκου
- /ðjaku/

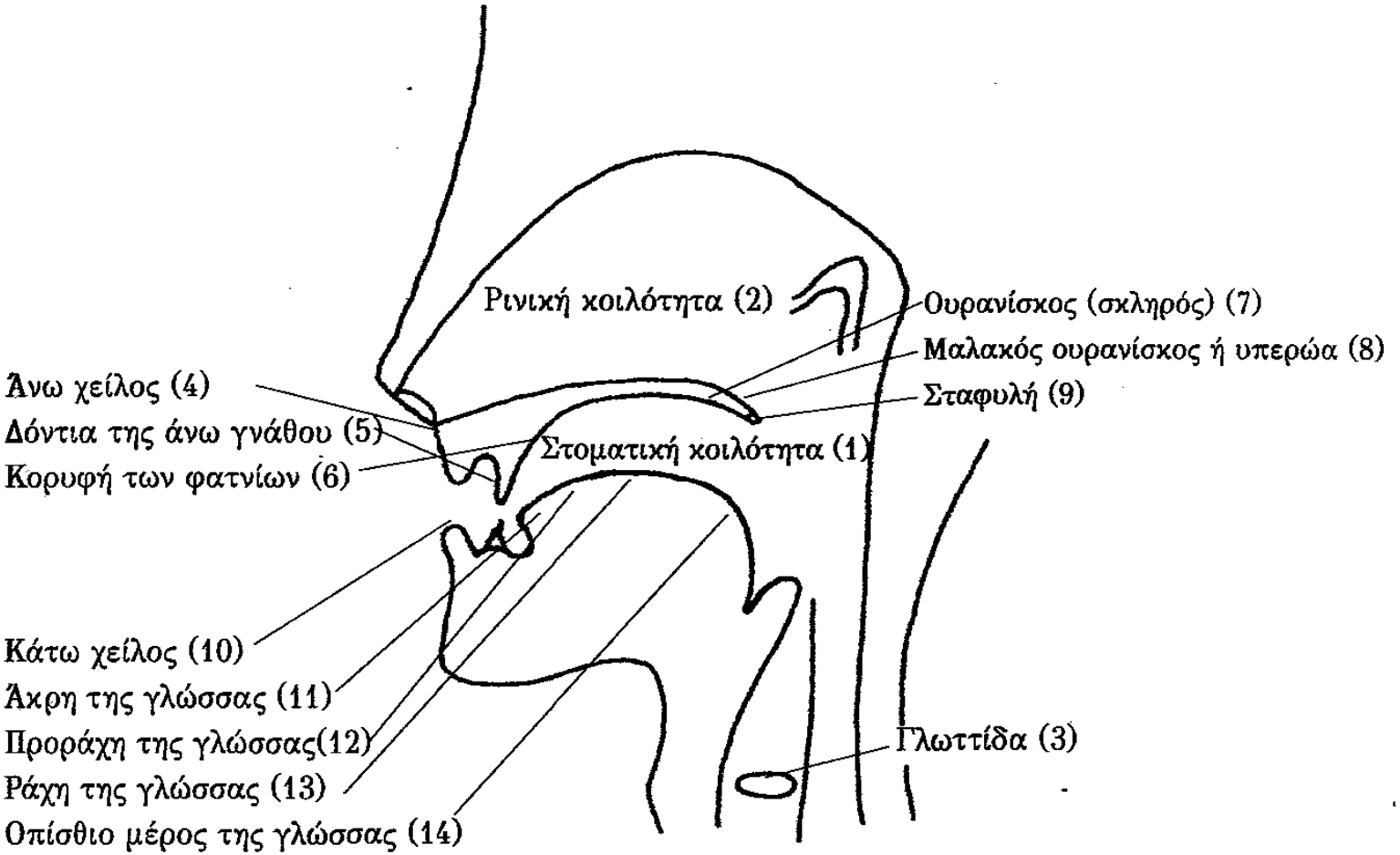
- Γ – γ – Ø – ɣ
- Approximate with grafo
- γράφω
- /ɣrafɔ/

- GR cap – GR low – ALB – IPA

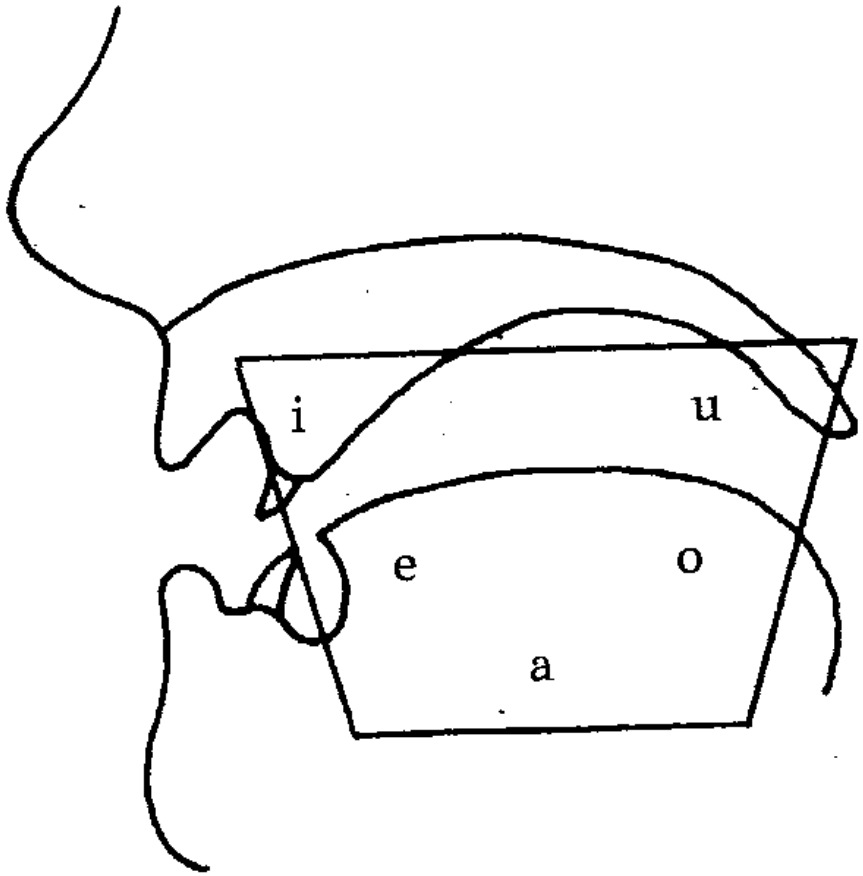
- ΟΥ – ου – U – u
- Pule
- Πούλε
- /pulɛ/

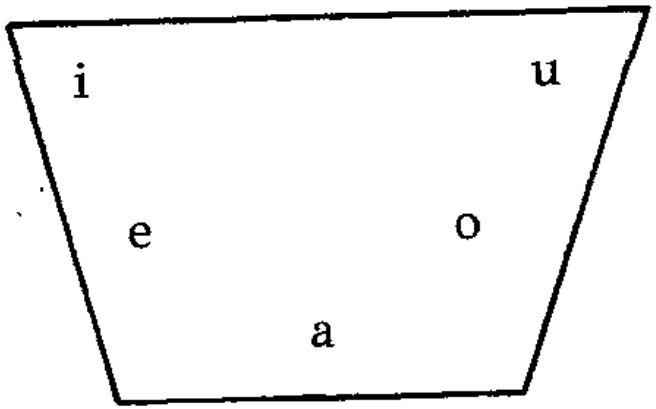
- Ø – Ø – Y – y
- ylber
- Approximate with Ιλμπερ
- /ylbɛɹ/

Organs of humans relevant with phonology [3]



Vowels of the Greek Language [4]





Phonetic Alphabets

- Why we need phonetic alphabets?
  - To be able to represent graphically all the phonemes exists in every human language
  - To be able to represent with the same symbol a single phoneme that is represented with different letters in different languages
  - To solve the restrictions of the written alphabets
    - γέρος (/ɟeros/)
    - γαρίδα (/ɣariða/)
- How many Exists ?
  - 2, IPA and SAMPA

IPA [5]

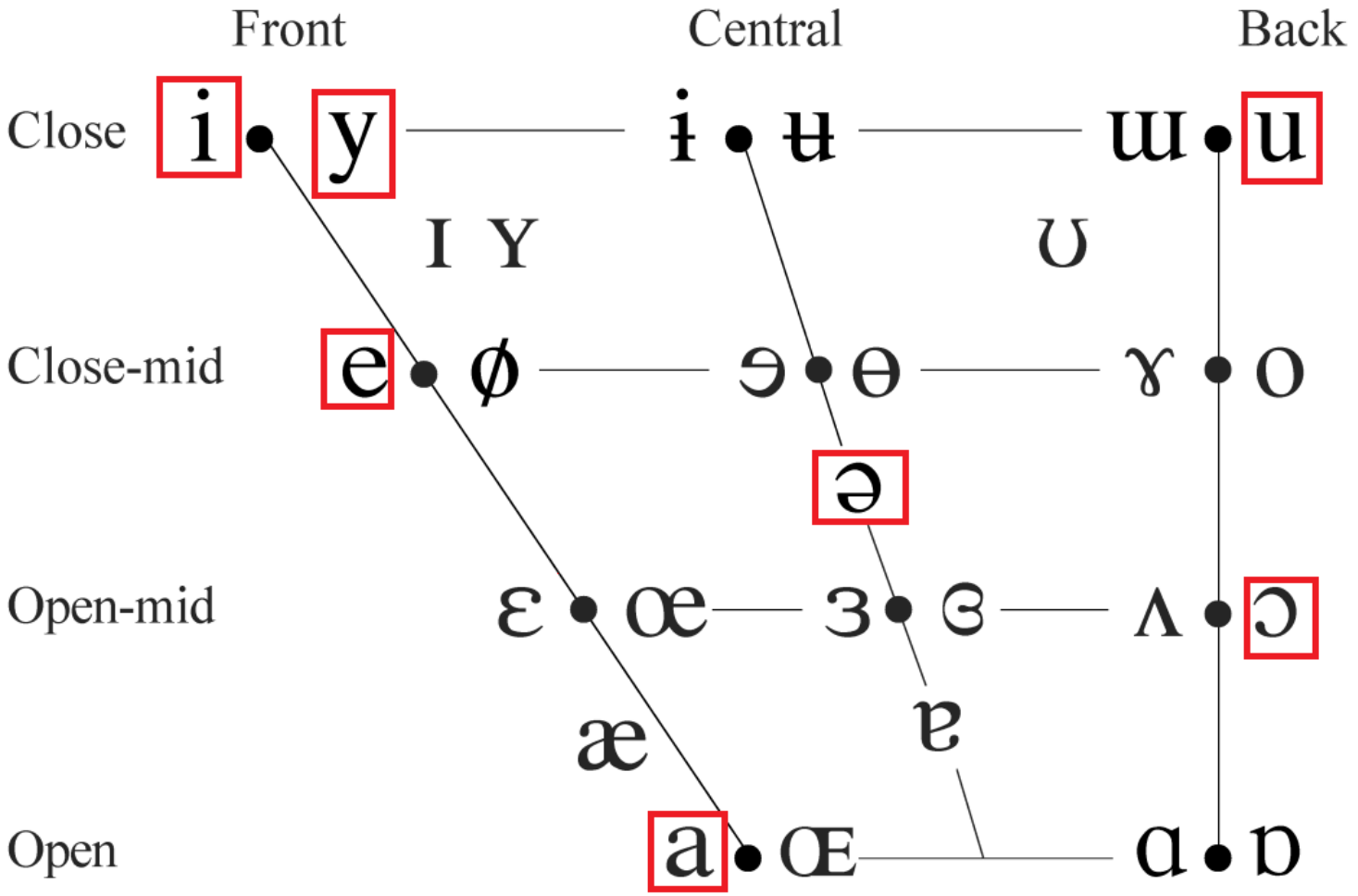
- The **International Phonetic Alphabet** (**IPA**) is an alphabetic system of phonetic notation based primarily on the Latin alphabet. It was devised by the International Phonetic Association in the late 19th century as a standardized representation of the sounds of spoken language.[1] The IPA is used by lexicographers, foreign language students and teachers, linguists, speech-language pathologists, singers, actors, constructed language creators and translators

More about IPA

- http://www.internationalphoneticalphabet.org/ipa-sounds/ipa-chart-with-sounds/ [6]
- https://www.internationalphoneticassociation.org/sites/default/files/IPA_Kiel_2015.pdf [7]
- https://www.internationalphoneticassociation.org/sites/default/files/IPA2005_3000px.png [8]

Vowels of the Albanian Language

A result of [8] and [9]



Where symbols appear in pairs, the one to the right represents a rounded vowel.

SAMPA [10]

- The **Speech Assessment Methods Phonetic Alphabet** (**SAMPA**) is a computer-readable phonetic script using 7-bit printable ASCII characters, based on the International Phonetic Alphabet (IPA).
- It was originally developed in the late 1980s for six European languages by the EEC ESPRIT information technology research and development program. As many symbols as possible have been taken over from the IPA; where this is not possible, other signs that are available are used, e.g. [@] for schwa (IPA [ə]), [2] for the vowel sound found in French *deux* (IPA [ø]), and [9] for the vowel sound found in French *neuf* (IPA [œ]).

More about SAMPA

- SAMPA at UCL
- http://www.phon.ucl.ac.uk/home/sampa/index.html [11]

- SAMPA for Greek
- http://www.phon.ucl.ac.uk/home/sampa/greek.htm [12]

Greek terms with IPA

- https://en.wiktionary.org/wiki/Category:Greek_terms_with_IPA_pronunciation [13]
- Today (30-4-2018) there are 3,481 items
- Items are organized per Greek letter: Αα Ββ Γγ Δδ Εε Ζζ Ηη Θθ Ιι Κκ Λλ Μμ Νν Ξξ Οο Ππ Ρρ Σσ Ττ Υυ Φφ Χχ Ψψ Ωω
- For each item there are:
  - Modern Greek: θάλασσα
  - Pronunciation with IPA: [ˈθalasa]
  - Usage examples
  - Declension
  - Related terms (some kind of thesaurus)

Albanian terms with IPA

- https://en.wiktionary.org/w/index.php?title=Category:Albanian_terms_with_IPA_pronunciation [14]
- Today (30-4-2018) there are 1,023 items
- Items are organized per Albanian letter: A B C Ç D Dh E Ë F G Gj H I J K L Ll M N Nj O P Q R Rr S Sh T Th U V X Xh Y Z Zh
- For each item there are:
  - Albanian word: ylber
  - Pronunciation with IPA: /ylbɛɹ/
  - English Translation: rainbow
  - Declension

Albanian Language Phonology [9]

- Albanian Language Phonology is simply
- Two tables of symbols are enough to transcribe from Albanian to IPA and vise versa

IPA	Description	Written as	English approximation
i	Close front unrounded vowel	i	seed
ɛ	Open-mid front unrounded vowel	e	bed
a	Open central unrounded vowel	a	father, Spanish casa
ə	Schwa	ë	about, the
ɔ	Open-mid back rounded vowel	o	law
y	Close front rounded vowel	y	French tu, German über
u	Close back rounded vowel	u	boot

Albanian Language Phonology – Consonants 1/3 [9]

IPA	Description	Written as	English approximation
m	Bilabial nasal	m	**m**an
n	Alveolar nasal	n	**n**ot
ɲ	Palatal nasal	nj	~**ni**on
ŋ	Velar nasal	ng	ba**ng**
p	Voiceless bilabial plosive	p	s**p**in
b	Voiced bilabial plosive	b	**b**at
t	Voiceless alveolar plosive	t	s**t**and
d	Voiced alveolar plosive	d	**d**ebt
k	Voiceless velar plosive	k	s**c**ar
g	Voiced velar plosive	g	**g**o

Albanian Language Phonology – Consonants 2/3 [9]

t͡s	Voiceless alveolar affricate	c	ha**ts**
d͡z	Voiced alveolar affricate	x	goo**ds**
t͡ʃ	Voiceless postalveolar affricate	ç	**ch**in
d͡ʒ	Voiced postalveolar affricate	xh	**j**et
c͡ç	Voiceless palatal affricate	q	~**ch**ina (RP)
ɟ͡ʝ	Voiced palatal affricate	gj	~**g**em (RP)
f	Voiceless labiodental fricative	f	**f**ar
v	Voiced labiodental fricative	v	**v**an
θ	Voiceless dental fricative	th	**th**in
ð	Voiced dental fricative	dh	**th**en

Albanian Language Phonology – Consonants 3/3 [9]

s	Voiceless alveolar fricative	s	**s**on
z	Voiced alveolar fricative	z	**z**ip
ʃ	Voiceless postalveolar fricative	sh	**sh**ow
ʒ	Voiced postalveolar fricative	zh	vi**s**ion
h	Voiceless glottal fricative	h	**h**at
r	Alveolar trill	rr	Spanish pe**rr**o
ɾ	Alveolar tap	r	Spanish pe**r**o
l	Alveolar lateral approximant	l	**l**ean
ɫ	Velarized alveolar lateral approximant	ll	ba**ll**
j	Palatal approximant	j	**y**es

Albanian Language Phonology
Not so simple - Not context free

- Before *q* and *gj,* the *n* is always pronounced /ɲ/ but it's not reflected in the orthography [9]. That means:
  …nq… → /… ɲcç…/ and not /…ncç…/
- Next example (from [9])
  ngjashëm → [nɟʑaʃəm]
  does not follows the tables.
  It should be transcribed [nɟaʃəm]

Do you pronounce the same way the letter on top in every following word

- **Q – q**
  Sufllaqe
  Xaxiq
  Qirici
  Qepë
  Qumeshtur
  Qentër

- **Ç – ç**
  çift
  çakmak

- **C – c**
  cigare

- **GJ – gj**
  Gjizë
  Gjumë
  Gjasht
  Gjermane
  Gjrokaster

- **XH – xh**
  Xhina
  Xhiola
  Maxhelaku
  Xheni

- **X – x**
  Xulja
  Xeni

Read for me

- Jaja
- Jala
- Jiala
- Jeros
- Jineka
- Jneka
- Jiatros
- Jatros
- Jliko
- juruni

Can you expain the differences

- Gjrokaster → Τζιροκαστερ
- Xhrokaster → Ντζίροκαστερ

Greek Language Phonology

- It is more difficult.
- There are a lot of context sensitive rules
- γέρος → /ʝeros/
  while
  γαρίδα → /ɣariða/
- αγγαρεία → /aŋgaria/
  αγγελία → /aŋʝelia/
  εγγόνι → /eŋgoni/
  έγγραφο → /eŋɣrafo/

Orthographic Transcription [15]

- **Orthographic transcription** is a transcription method that employs the standard spelling system of each target language.
- Examples of orthographic transcription are "Pushkin" and "Pouchkine", respectively the English and French orthographic transcriptions of the surname "Пу́шкин" in the name Алекса́ндр Пу́шкин (Alexander Pushkin).
- Thus, each target language (English and French) transcribes the surname according to its own orthography.

Elaborating Orthographic Transcription

We can form training sets or corpus like:

Albanian word (1)	Orthographic Transcription Greek (2)	Phonetic Transcription (IPA) (3)	Translation Greek (4)
Bagazh	μπαγκάζ	[bagɑʒ]	βαλίτσα
Dashuroj	νταssσουρόι	[daʃurɔj]	αγαπάω
dymbëdhjetë	ντουμπαδιέτ(α)	[dymbə'ðjɛt(ə)]	Δώδεκα
Derr	ντερ	dɛr	γουρούνι
Buzëqeshje	μπουζατσέσχιε	/buzə'cɛʃjɛ/, [bus'cɕɛʃjɜ]	γελώ
Kolloface	κολοφάτσε	[kɔɬɔˈfatsɛ]	λουκάνικο
Kuptoj	κουπτόι	[kup'tɔj]	καταλαβαίνω
Gjumë	τζιούμ(α)	/ɟumə/ /dʑumə/	κοιμάμαι

Problems we can solve

For a Greek User

Usage	Input	Output
A	(4) Γουρού νι	(1), (2) derr, ντερ
B	(1) Derr	(4), (2) γουρούνι, ντερ

We assume that a simple translation dictionary exist that offers:
(4) → 1 and (1) → (4).

How A can be done:
1st step: (4) → (1); from translation dictionary
2nd step: (1) → (2); can already exist (in training set),
Otherwise (1) → (3) → (2); two steps are needed

So we need Transcribers for **Albanian to IPA** and **IPA to Greek**

Problems we can solve

Similar usages can supported for the Albanese User

Usage	Input	Output
C	kolloface	λουκάνικο, lukaniko
D	λουκάνικο	kolloface, lukaniko

We need Transcribers for **Greek to IPA** and **IPA to Albanian**

Machine Learning of Phonological Rules for Greek Transcription

- Is it possible to create some program that learns how to transcribe from Greek to IPA and from IPA to Greek?
- Yes, We think so,
- We need some resources, like “Wiktionary, Greek terms with IPA pronunciation” [13] and some Algorithm.
- We start with Greek, because it has many rules (dependency of contexts) for phonological transcription.
- The algorithm should uncover (mechanically learn, mine) these rules.

First step

- Consider words having the same number of Greek letters and IPA symbols in the transcription
- Example: πορτοκάλι → /portokali/
- With the one by one correspondence we can conclude:
  - π transcribe to p
  - ο transcribe to o
  - ρ transcribe to r
  - …

Fist Step after many words

α={ a=323, k=4, ɐ=2, n=1, s=1, e=1 },	β={ v=67 },	γ={ ɣ=51, ɟ=20, ŋ=5, ɣ=1, g=1, ɬ=1 },

First step erroneous results

- Can have some erroneous results
- Example:
  εγκέφαλος → /eŋgefalos/
  conclude: γ transcribe to ŋ
  it is on of the 5 cases (out of 79) we have found where γ transcribe to ŋ
- Another example:
  ευθανασία → /efθanasia/
  conclude υ transcribe to f
  it is one of the 9 cases (out of 58) we have found where υ transcribe to f

Step two

- We keep only transcriptions having a percentage above a predefined treshold
- For 20% transcription table is reduced to:

α={ a=323, },	β={ v=67 },	γ={ ɣ=51, ɟ=20, },

Third Step

- Consider words having one more Greek letter than the symbols in the IPA transcription
- Example: ουρανός → /uranos/
- With the one by one correspondence and respecting the results of the second step, the algorithm can conclude:
  - ου transcribes to u
  - ει transcribes to i
  - οι transcribes to i

Third Step after many words

ου={ u=1, },	ει={ i=1, },	οι={ i=1, },	αι ={ e=1, ε=1, },

Fourth Step

- Consider words having one less Greek letter than the symbols in the IPA transcription
- Example: ψάρια → /psaria/
- With the one by one correspondence and respecting the results of the second step, the algorithm can conclude:
  - ψ transcribes to ps
  - ξ transcribes to ks

Fourt Step after many words

ξ={ ks=10, },	ψ={ ps=15, },

Fifth Step

- Consider words having not resolved in previous steps
- GR letters can be +1 | +2 | -1 | -2 relatively to IPA symbols in transcription
- Example are:
  - ηλιόλουστος → /iʎɔlustɔs/
  - θηλιά → θiʎa
  - γιάννα → /ʝana/
  - γιαούρτι → /ʝaurti/
  - γράμμα → /ɣrama/
  - γέννηση → /ʝɛnisi/
  - μελισσοκομία → /melisokomia/
- With the one by one correspondence and respecting the results of all previous steps, the algorithm can conclude interesting valid transcriptions:

Fifth Step after many words

λι={ Λ=2, },	μμ={ m=1, },	νν={ n=2, },	σσ={ s=1, },	γι={ ɟ=2, },

Protected couples

- There are couples of letters that correspond phonetically to IPA couples of symbols.
- It is wrong to split the couple and consider each letter separately.
- These couples sometimes are also depended to their context (usually previous and next letter).
- These letters should be examined by the next step. For this reason, the operator of the Algorithm should have declare these couples in order the words having them not to be considered by previous steps.
- Such couples we call protected
- For the Greek language we suggest:
  γγ, γκ, τσ, τζ, μπ, ντ, αυ, ευ
- Also the some stand for some triangles:
  ντσ. ντζ

Sixth Step

- The Algorithm considers only words not resoled by previous steps.
- It tries to find correspondences respecting all the previous findings and resolving the protected couples (and triangles).
- Given:
  - καλιαρντά → /kaʎarda/
  - Μέτσοβο → /metsovo/
  - τσιμπούκι → /tsimbuci/
  - μπαμπάς → /babas/
  - …
  - Εύβοια → /evia/
  - Ευγενία → /evʝenia/
  - ευθανασία → /efθanasia/
  - αυγό → /avɣo/
  - αυτοκίνητο → /aftɔcinitɔ/

Sixth Step after many words

ντ={ d=1, },	τσ={ ts=1, },	μπ={ mb=2, b=2, },	τζ={ ts=1, dz=2, },	γγ={ ŋg=2, ŋɟ=1, ŋɣ=1, },	γκ={ ŋg=2, ŋɟ=1, g=2, ɟ=1, },	ɛυ={ e=1, ev=1, ef=1, },

Contextual data

- The unification of all previous tables (step 2 to 6) are the set of transcription rules
- However, there are ambiguity cases. For example when the Greek word contains κ when it is transcribed to k and when it is transcribed to c?
- The algorithm should also learn disambiguation of rule usage.
- To do this, the algorithm should keep the contexts.
- Next we see Greek couple αυ with contextual data:

```
αυ={
   av=1, -αυγ
   af=1, -αυτ
},
```

- with more data can become:
  αυ={
    av=8, -αυγ, -αυλ, -αυν, μαυρ, ταυρ, ραυλ, βαυα, καυλ,
    af=4, -αυτ, εαυτ, ναυπ,
  },
- Can be generalized to:
  αυ={
    av=8, when next letter is one of γ, λ, ν, ρ, α,
    af=4, when next letter is one of τ, π,
  },
- And with more data can be generalized to:
  αυ={
    av=x_1, when next leter is vowel or voiced consonant,
    af=x_2, when next letter is voiceless consonant,
  },

Other rules we expect to mine

- μπ={
  mb=r_1, after vowel
  b=r_2, in the beginning of word
  or after consonant
  },
- νι={
  ɲ=t_1, when next letter is vowel
  otherwise ni
  },

References

- [1] Wikipedia, Phonology https://en.wikipedia.org/wiki/Phonology
- [2] Wikipedia, Phoneme ttps://en.wikipedia.org/wiki/Phoneme
- [3] Marina Nespor, ΦΩΝΟΛΟΓΙΑ, Chapter 2, Schema 2, ISBN: 960-378-083-9
- [4] Marina Nespor, ΦΩΝΟΛΟΓΙΑ, Chapter 2, Schema 3, ISBN: 960-378-083-9
- [5] Wikipedia, International Phonetic Alphabet https://en.wikipedia.org/wiki/International_Phonetic_Alphabet

References

- [6] IPA chart with sounds http://www.internationalphoneticalphabet.org/ipa-sounds/ipa-chart-with-sounds/
- [7] IPA revised 2015 https://www.internationalphoneticassociation.org/sites/default/files/IPA_Kiel_2015.pdf
- [8] IPA 2005 https://www.internationalphoneticassociation.org/sites/default/files/IPA2005_3000px.png
- [9] Albanian Language Phonology https://en.wikipedia.org/wiki/Albanian_language#Phonology
- [10] Wikipedia SAMPA https://en.wikipedia.org/wiki/SAMPA

References

- [11] SAMPA http://www.phon.ucl.ac.uk/home/sampa/index.html
- [12] SAMPA for Greek http://www.phon.ucl.ac.uk/home/sampa/greek.htm
- [13] Wiktionary, Greek terms with IPA pronunciation https://en.wiktionary.org/wiki/Category:Greek_terms_with_IPA_pronunciation
- [14] Wiktionary, Albanian terms with IPA pronunciation https://en.wiktionary.org/wiki/Category:Albanian_terms_with_IPA_pronunciation
- [15] Wikipedia, Orthographic Transcription https://en.wikipedia.org/wiki/Orthographic_transcription