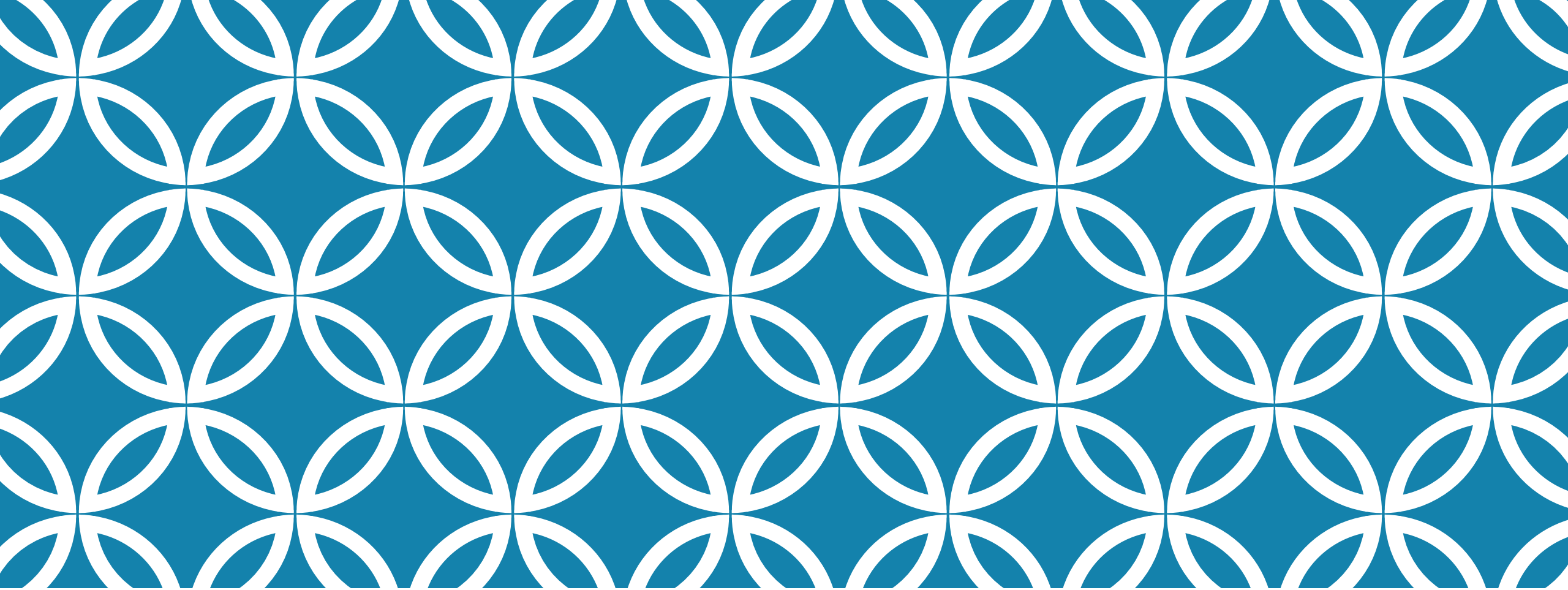


DESIGNING A MODERN GREEK SENTENCE CORPUS FOR AUDIOLOGICAL AND SPEECH TECHNOLOGY RESEARCH

ANNA SFAKIANAKI
UNIVERSITY OF CRETE
asfakianaki@csd.uoc.gr

MOTIVATION

- speech audiometry
  - hearing assessment
  - central auditory processing disorders

Indiamart.com

- speech technology
  - mobile communications
  - speech perception in noise
  - development of algorithms to enhance intelligibility

- Need for specially designed corpora

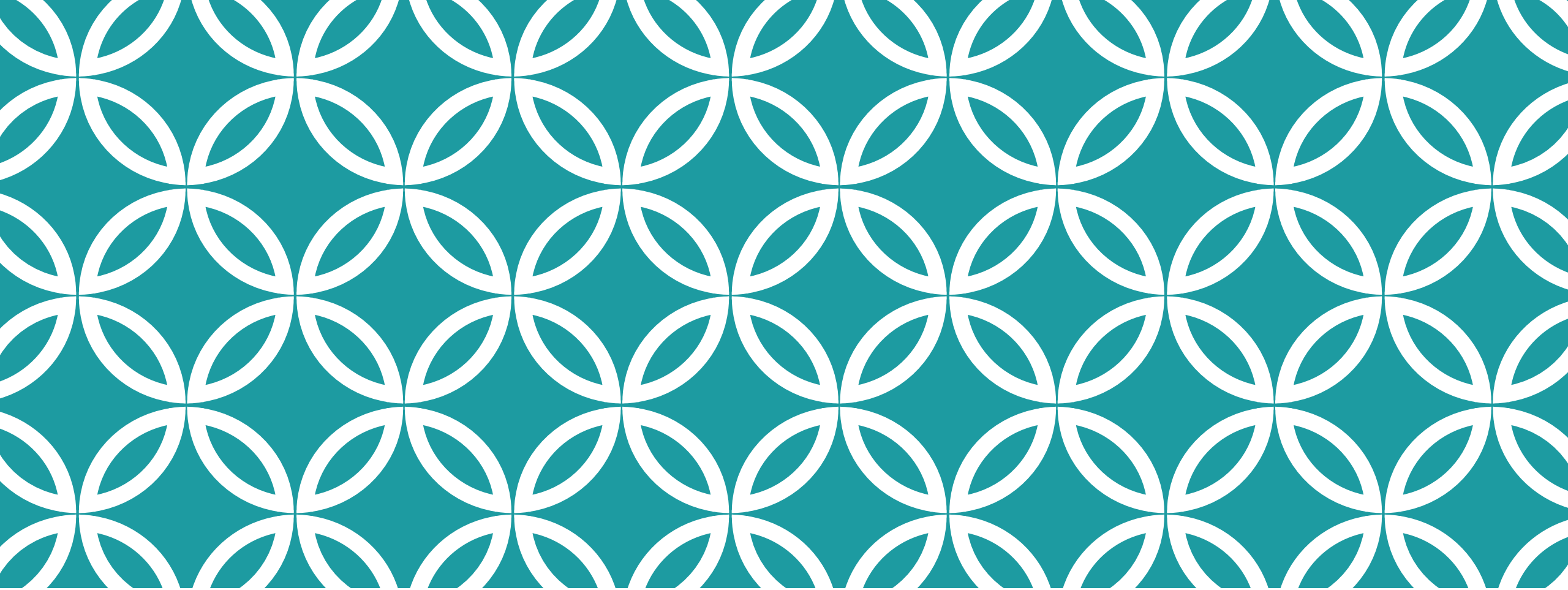
MOTIVATION

Horizon 2020

http://www.enrich-etn.eu/

ENRICH EUROPEAN TRAINING NETWORK

- Network: 8 universities & 6 industry & clinical partners training 14 PhDs
- Improve **hearing aids** and **speech synthesis** for individuals with reduced capacity in listening/speaking
- *The fundamental objective of the ENRICH network is to* ***modify or augment speech*** *with additional information to* ***make it easier to process***.
- Listening tests to investigate relationship between forms of natural and synthetic speech and cognitive effort



BACKGROUND

BACKGROUND

- Existing material in Modern Greek
  - 3 50-word lists of bisyllabic real words
    (Iliadou, Fourakis, Vakalos, Hawks & Kaprinis, 2006)
  - 4 50-word lists of bisyllabic real words
    (Trimmis, Papadeas, Papadas, Naxakis, Papathanasopoulos & Goumas, 2006)
  - 2 50-word lists of nonsense monosyllables with possible CV, VC, and CVC phonemic combinations (Trimmis, Vrettakos, Gouma, Papadas, 2012)
  - 5 50-word lists of bisyllabic nonsense words
    (Trimmis, Mourtzouchos, Naxakis, Papadas, & Goumas, 2013)

- No sentence material available for such purposes.

- Natural conversation
- Word transitions, reductions, intonation
- Redundancy
- Semantic & syntactic cues
- Noise-reduction algorithms

BACKGROUND

- Sentence corpora constructed in other languages using various methods and tools
  - **English**
    - SPIN Test (Speech perception in noise) (Kalikow *et al.*, 1977)

    8 lists of 50 sentences where the keyword is the final word, a monosyllabic noun of either high or low predictability
    - Examples

    (HP) The watchdog gave a warning **growl**.

    (LP) The old man discussed the **dive**.

BACKGROUND

- Sentence corpora constructed in other languages using various methods and tools
  - **English**
    - HINT Test (Hearing in noise) (Nilsson et al., 1994)

    25 lists of 10 sentences designed for sSRTs, where all content words are scored. Based on BKB sentences for use with British children (Bench & Bamford, 1979)
    - Examples

    A **boy fell** from the **window**.

    The **wife helped** her **husband**.

    Big dogs can be **dangerous**.

BACKGROUND

- **Spanish**
  - Test similar to SPIN Test (Kalikow *et al.*, 1977)
    6 lists of 25 HP sentences and 6 lists of 25 LP sentences
    (Cervera & González-Alvarez, 2010)

- **Swedish, Danish & German**
  - Hagerman-type test (Hagerman, 1982)
  - Oldenburg sentence test (OLSA) (Wagener *et al.*, 1999)
    5 x 10 word matrix with columns containing 10 names, verbs, numbers, adjectives & objects (LP)

- **Dutch**
  - Automated selection of 35,000 sentences (up to 9 syllables long) from large databases and manual selection of 1311 sentences
    (Versfeld *et al.*, 2000)

BACKGROUND

Name	*Verb*	*Number*	*Adjectiv*	*Object*
Peter	bekommt	drei	große	Blumen
Peter	*gets*	*three*	*big*	*Flowers*
Kerstin	sieht	neun	kleine	Tassen
Kerstin	*sees*	*nine*	*little*	*cups*
Tanja	**kauft**	sieben	alte	Autos
Tanja	***buys***	*seven*	*old*	*cars*
Ulrich	gibt	acht	nasse	Bilder
Ulrich	*gives*	*eight*	*wet*	*pictures*
Britta	schenkt	vier	schwere	**Dosen**
Britta	*donates*	*four*	*heavy*	***cans***
Wolfgang	verleiht	fünf	grüne	Sessel
Wolfgang	*lends*	*five*	*green*	*chairs*
Stefan	hat	zwei	teure	Messer
Stefan	*has*	*two*	*expensive*	*knives*
Thomas	gewann	achtzehn	**schöne**	Schuhe
Thomas	*wons*	*eighteen*	***pretty***	*shoes*
Doris	nahm	**zwölf**	rote	Steine
Doris	*took*	***twelve***	*red*	*stones*
Nina	malt	elf	weiße	Ringe
Nina	*paints*	*eleven*	*white*	*rings*

(Wagener *et al.*, 1999:72), Table 1

BACKGROUND

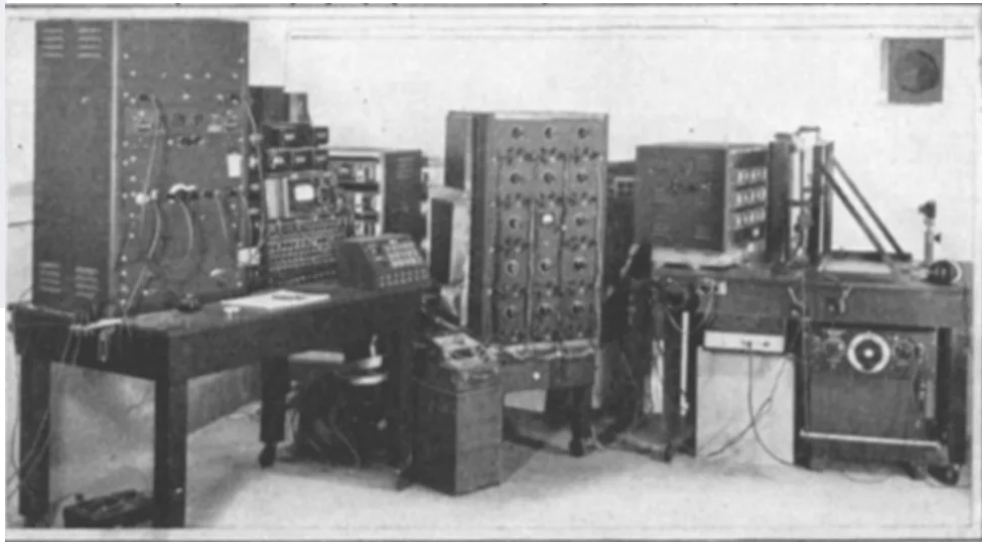
- **Finnish & English**
  - Simultaneous development of material in English and Finnish.
    1000 sentences extracted from text corpora (9-12 syllables for Finnish & 7-9 for English) and balanced for word frequency and phone distribution (Vainio *et al.*, 2005)

- **Thai**
  - 313 sentences from 89 children's stories (familiarity). Tool development for word replacement to achieve PB. Extra step for predictability.
    (Munthuli *et al.*, 2015)

BACKGROUND

Faculty & Staff of Harvard Psycho-Acoustic Laboratory, Memorial Hall (1945)

Founder: S.S. Stevens



Source: https://gizmodo.com/the-harvard-sentences-secretly-shaped-the-development-1689793568

BACKGROUND

- **Harvard Corpus** (Rothauser *et al.*, 1969)
  - 72 PB lists
  - 10 sentences in each list
  - 5 keywords in each sentence
  - 1-2 syllable-words
- Recommended for speech quality measurements by IEEE (Institute of Electrical and Electronics Engineers)
- Has been used extensively in speech intelligibility testing (e.g. Bradlow *et al.*, 1996; Hu & Loizou, 2010; Cooke *et al.*, 2013)
- **S-Harvard Corpus** recently constructed for Spanish (Aubanel et al., 2014)
- **Gr-Harvard Corpus…**

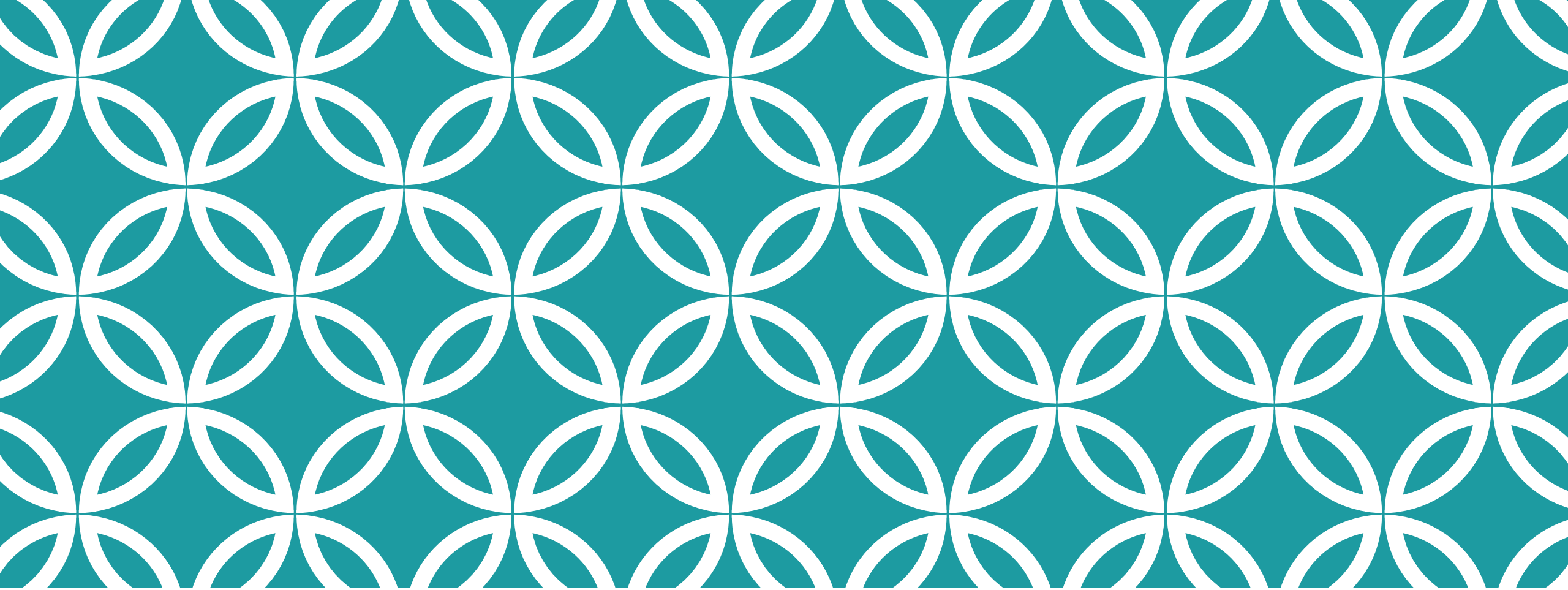
1965 Revised List of Phonetically Balanced Sentences (Harvard Sentences)

List 1

1. The birch canoe slid on the smooth planks.
2. Glue the sheet to the dark blue background.
3. It's easy to tell the depth of a well.
4. These days a chicken leg is a rare dish.
5. Rice is often served in round bowls.
6. The juice of lemons makes fine punch.
7. The box was thrown beside the parked truck.
8. The hogs were fed chopped corn and garbage.
9. Four hours of steady work faced us.
10. A large size in stockings is hard to sell.

"These materials have been the gold standard."

David Pisoni
Director of Speech Research Laboratory
Indiana University



GR-HARVARD CORPUS DESIGN

CORPUS DESIGN CRITERIA

- Each sentence must have
  - 5-9 words in total
  - **5 keywords**
- Each word must be
  - **1-3 syllables** long
  - repeated as little as possible throughout the corpus
- Keywords must be combined so that sentences are **meaningful and semi-predictable**
- Avoidance of proverbs and stereotyped phrases

- Keywords: Mostly content words
- Non-keywords: pronouns, articles, auxiliary verbs (e.g. είμαι, έχω), function words (e.g. αν, ας, ή)
- Function words that are marked as keywords
  - δεν, μην, σαν, πιο, πια, προς, και (emphatic), άμα, τι (emphatic/exclamatory), ενώ, αντί, μπρος, όσα, αλλά
- What is considered as keyword also depends on sentence structure and meaning.
  - "Και" is not marked as a keyword, unless it is emphatic.

CORPUS DESIGN

- Inspiration from
  - Original Harvard sentences
  - Greeklex 2
  - Greek online dictionaries
  - Internet (google search)

Greeklex 2 (Kyparissiadis et al., 2017)
A comprehensive lexical database with part-of-speech, syllabic, phonological, and stress information as well as several measurements of word similarity and phonetic information

- The **wide road shimmered** in the **hot sun**.
- Ο **φαρδύς δρόμος λάμπει** στον **καυτό ήλιο**.
  [The wide road shines in the hot sun.]

- **Let** it **burn**, it **gives** us **warmth** and **comfort**.
- Η **σόμπα** στη **γωνιά παρέχει ζεστασιά** και **θαλπωρή**.
  [The heater in the corner provides warmth and cosiness.]

- A **siege will crack** the **strong defense**.
- **Παρά** την **ισχυρή άμυνα**, τους **νίκησε** ο **εχθρός**.
  [Despite the strong defense, the enemy defeated them.]

CORPUS DESIGN

- Inspiration from
  - Original Harvard sentences
  - Greeklex 2
  - Greek online dictionaries
  - Internet (google search)

- The **man wore** a **feather** in his **felt hat**.
- Το **μουσείο εκθέτει τοπικές φορεσιές** και **όπλα**.
  [Local costumes and arms are exhibited at the museum.]

- The **pods** of **peas ferment** in **bare fields**.
- **Άφησε λίγο** τις **γαρίδες** να **βράσουν** στο **ζουμί** τους.
  [(He/she) let the shrimp boil in its own juices for a while.]

- **These days** a **chicken leg** is a **rare dish**.
- Ο **κόσμος τρώει τακτικά ψητό κρέας**.
  [People have roast meat regularly.]

CORPUS DESIGN

- Phonetic transcription
  - manual
  - in SAMPA
  - cross-checked with **IPLR** (Protopapas et al., 2010) & GreekLex2
  - No sandhi at word boundaries
  - All cases of nasal consonants followed by homorganic stops (e.g. [mb],[nd]) were simplified by dropping the nasal as in IPLR & GreekLex2

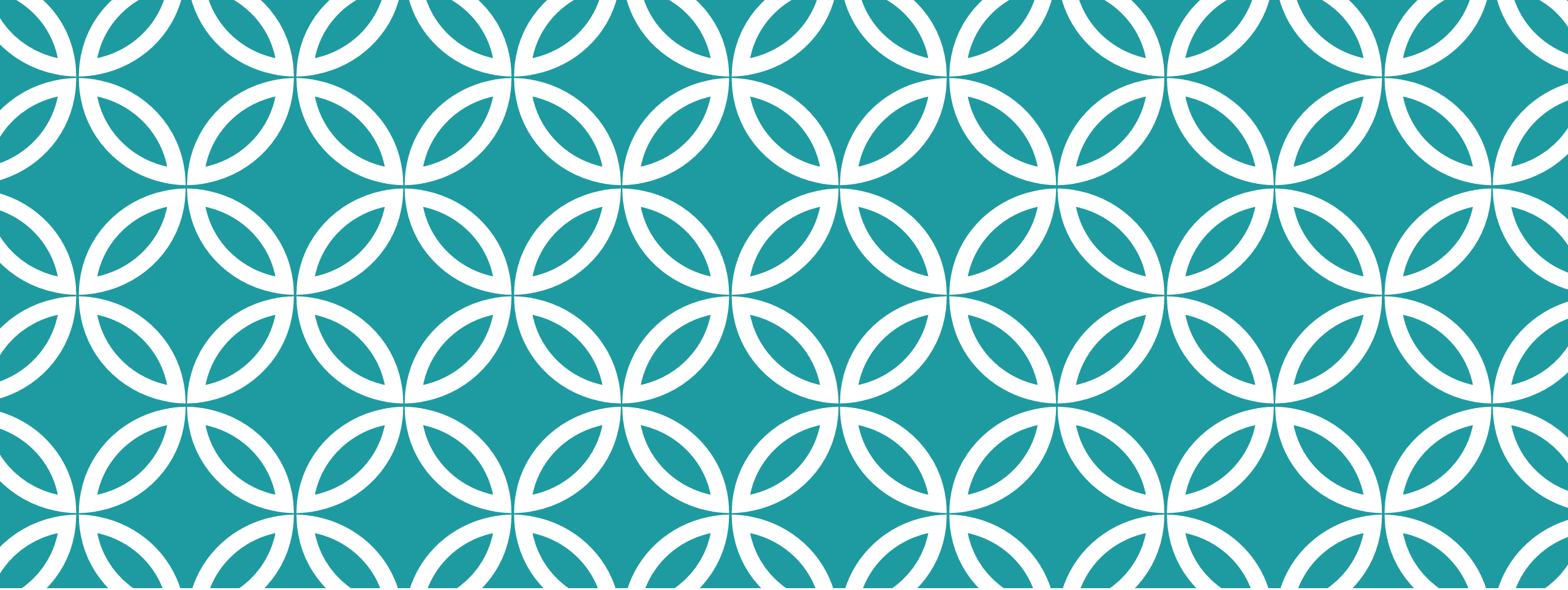
Online psycholinguistic resource for Greek based on analyses of written corpora combined with text processing technologies developed at the ILSP.
http://speech.ilsp.gr/iplr

Examples

- Το **σχέδιο δράσης** είναι **ασαφές προς** το **παρόν**.
  [The plan of action is unclear at the moment.]

  to."sCeDio."Drasis."ine.asa"fes.pros.to.pa"ron
- Ο **φαρδύς δρόμος λάμπει** στον **καυτό ήλιο**.
  [The wide road shines in the hot sun.]

  o.far"Dis."Dromos."labi.ston.ka"fto."iLo

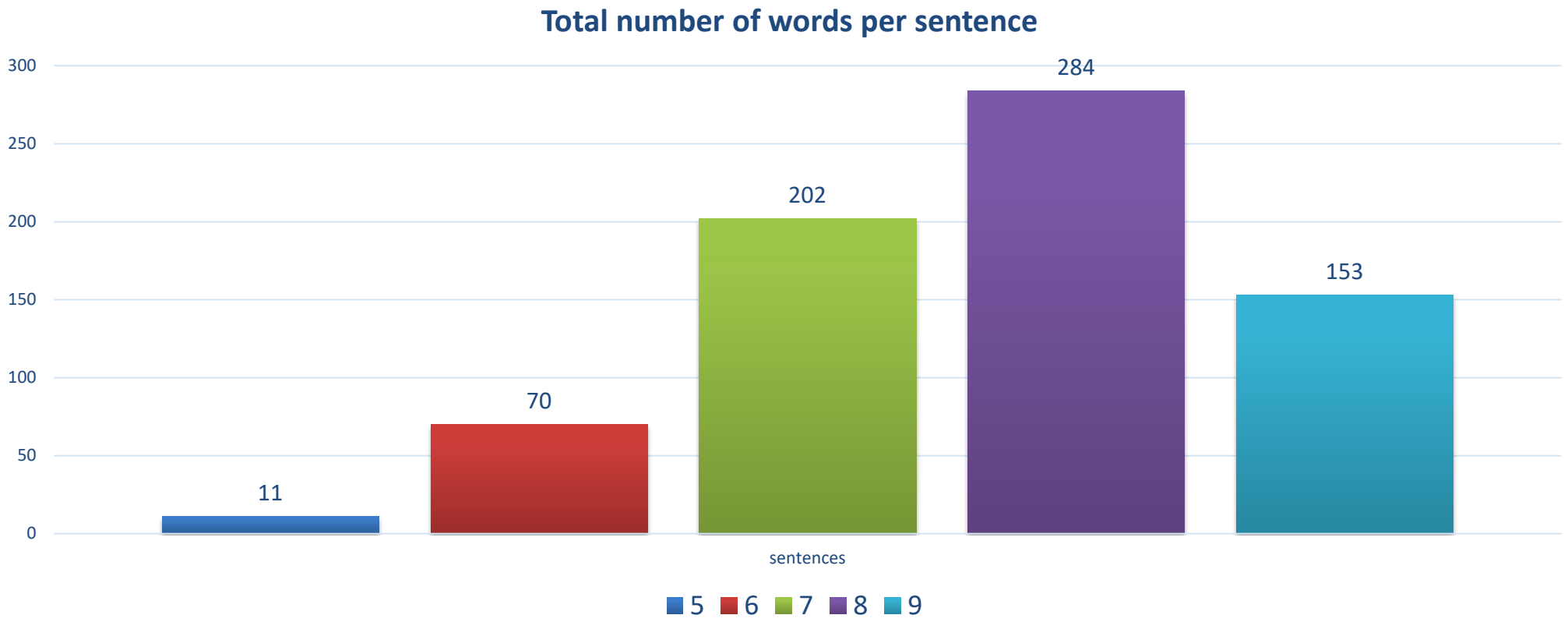


GR-HARVARD CORPUS STATISTICS

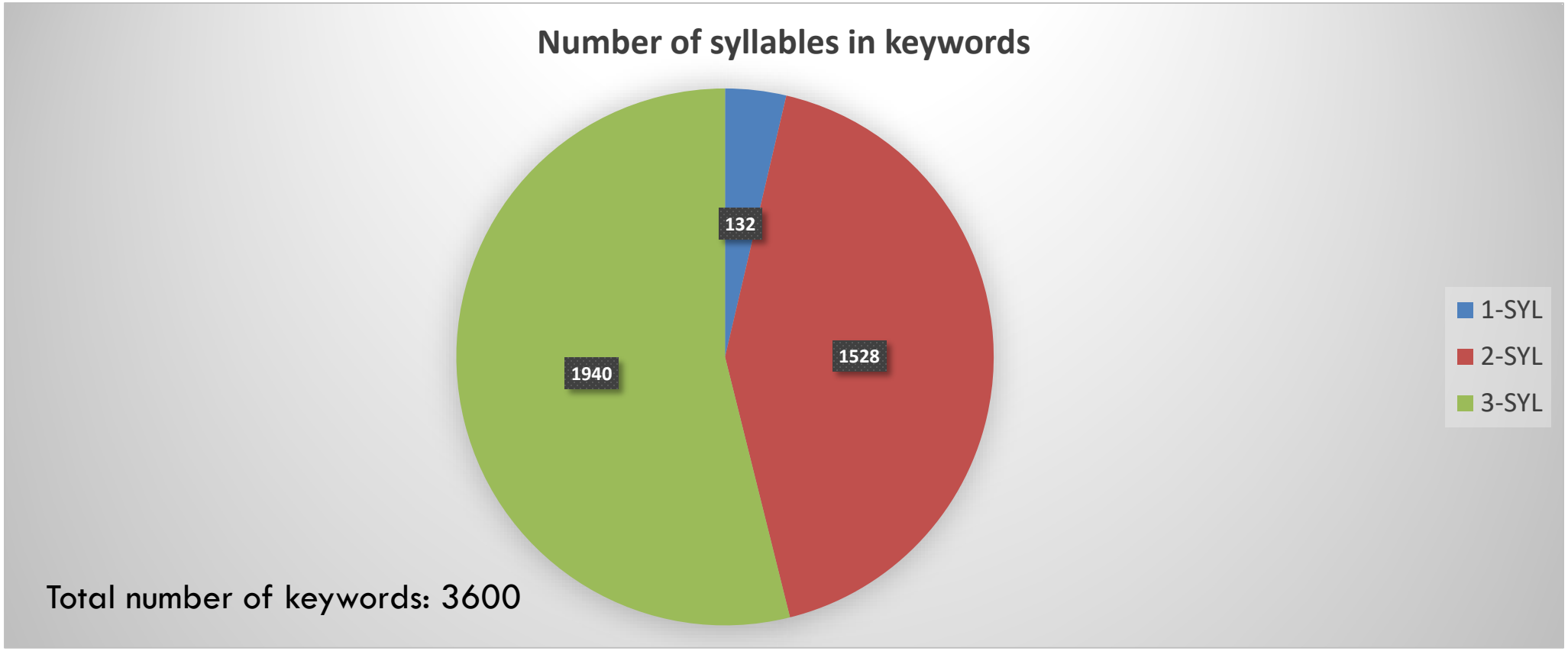
CORPUS STATISTICS

- 720 sentences
- 5 keywords per sentence
- 3,600 keywords in total
- Total number of words per sentence: 5 – 9
- Examples
  - 5 words → **Ψύχοντας νερό φτιάχνεις καθαρό πάγο**.
  - 6 words → **Βρέθηκε νέο φάρμακο κατά** του **διαβήτη**.
  - 7 words → **Εννιά εργάτες σκάβουν** τον **τόπο** για **αρχαία**.
  - 8 words → **Χώμα** και **σκόνη έτσουξαν** τα **μάτια** του **κοριτσιού**.
  - 9 words → Ο **τολμηρός λοχίας σύρθηκε** στο **πεδίο** με τις **νάρκες**.

TOTAL NUMBER OF WORDS PER SENTENCE

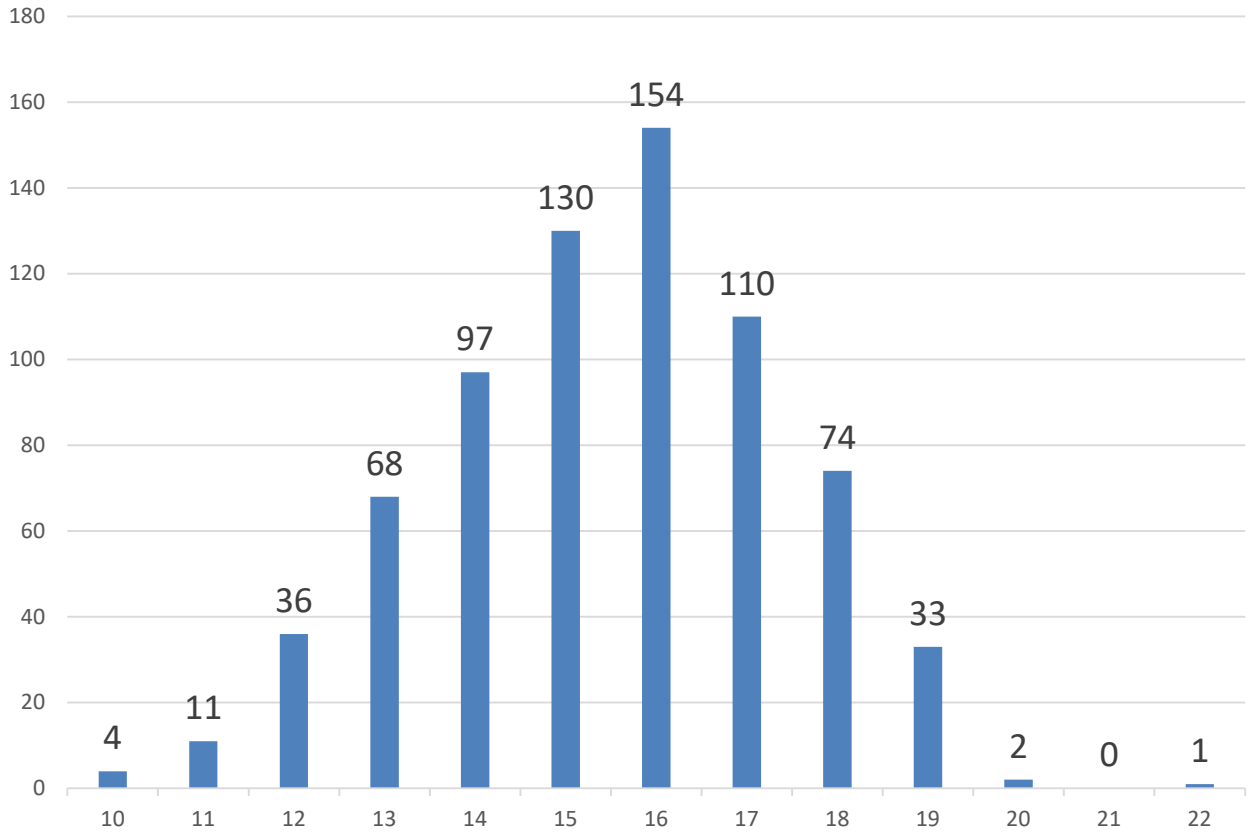


NUMBER OF SYLLABLES IN KEYWORDS



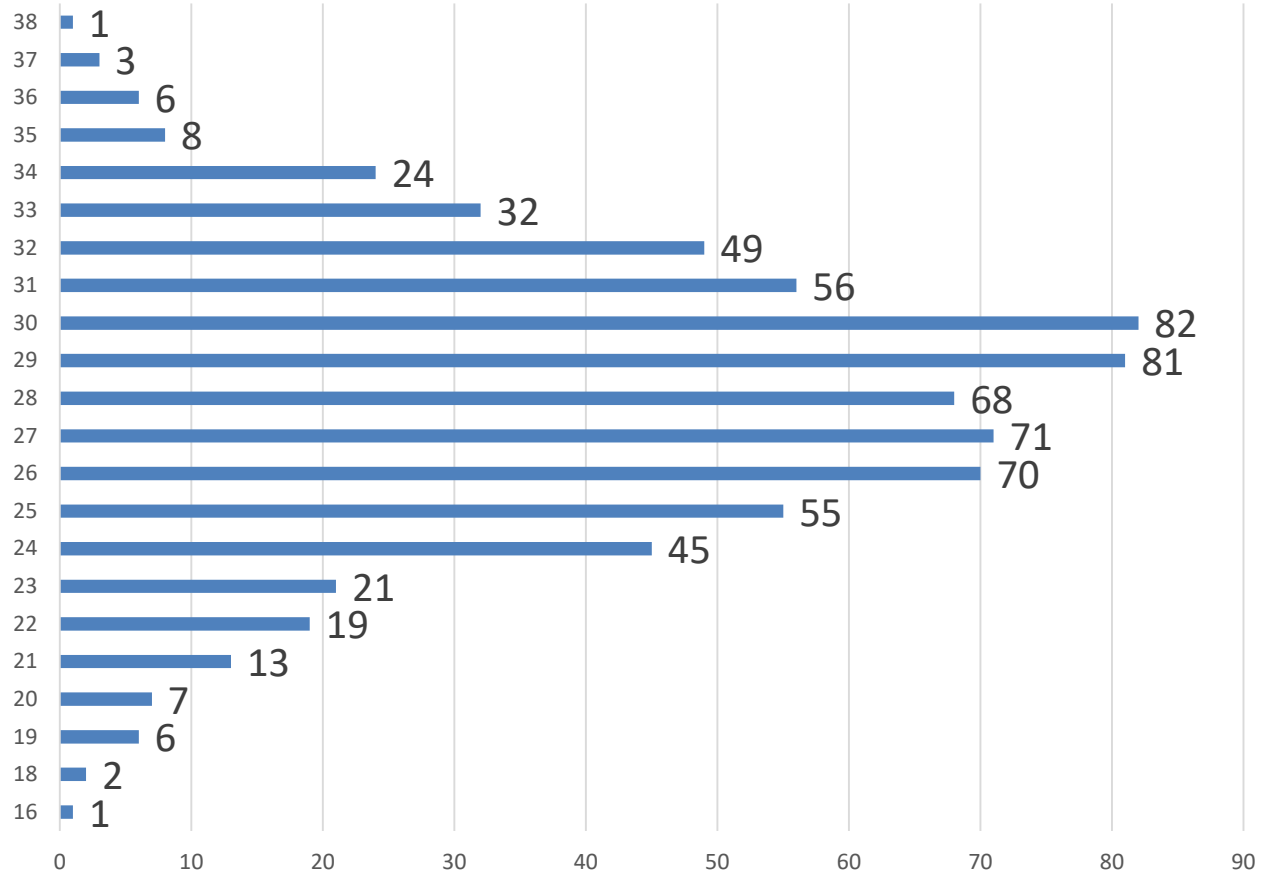
TOTAL NUMBER OF SYLLABLES PER SENTENCE

- 10 to 22 total number of syllables per sentence



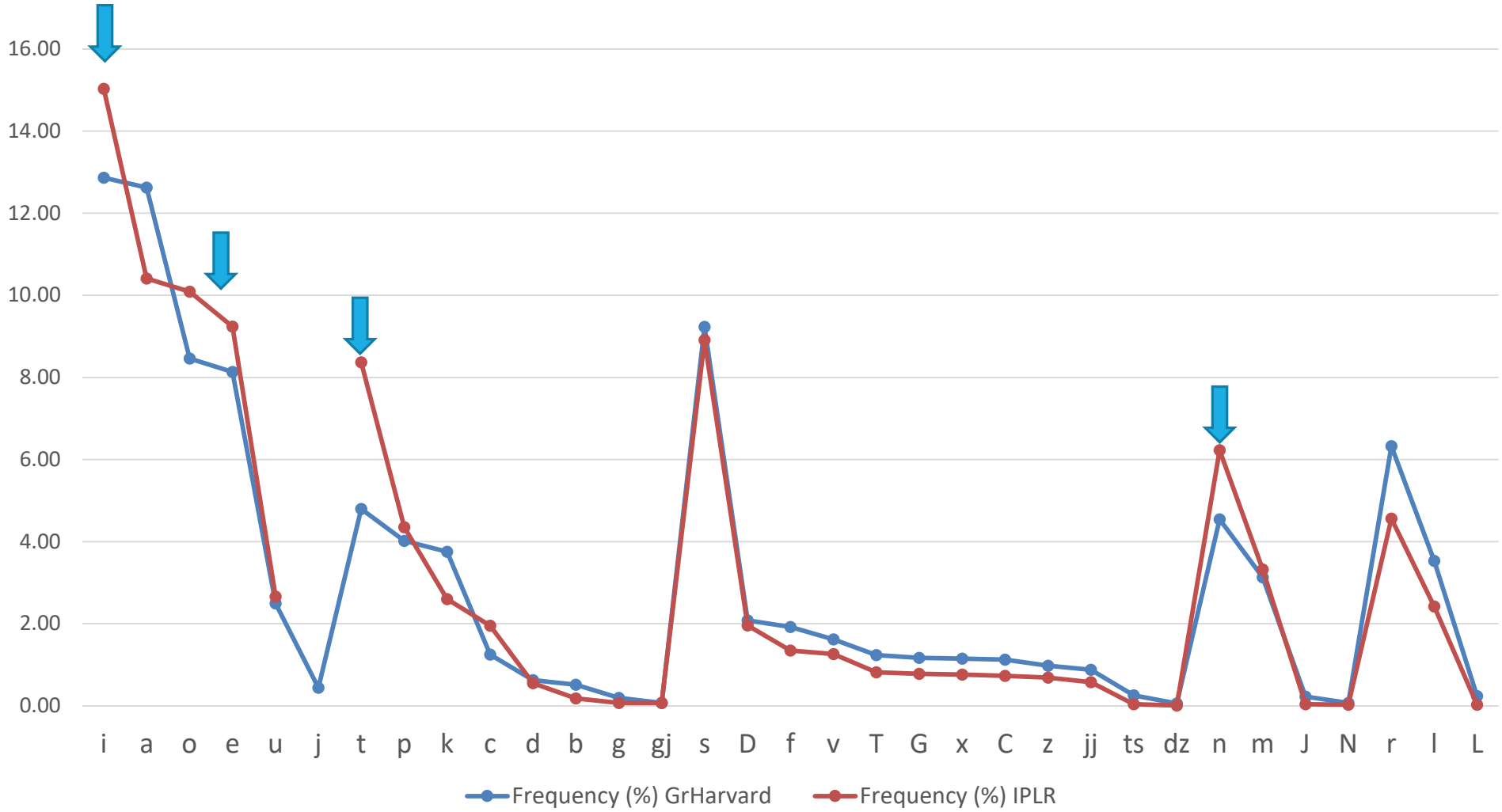
NUMBER OF PHONEMES IN KEYWORDS

- 16 to 38 total number of phonemes in keywords per sentence



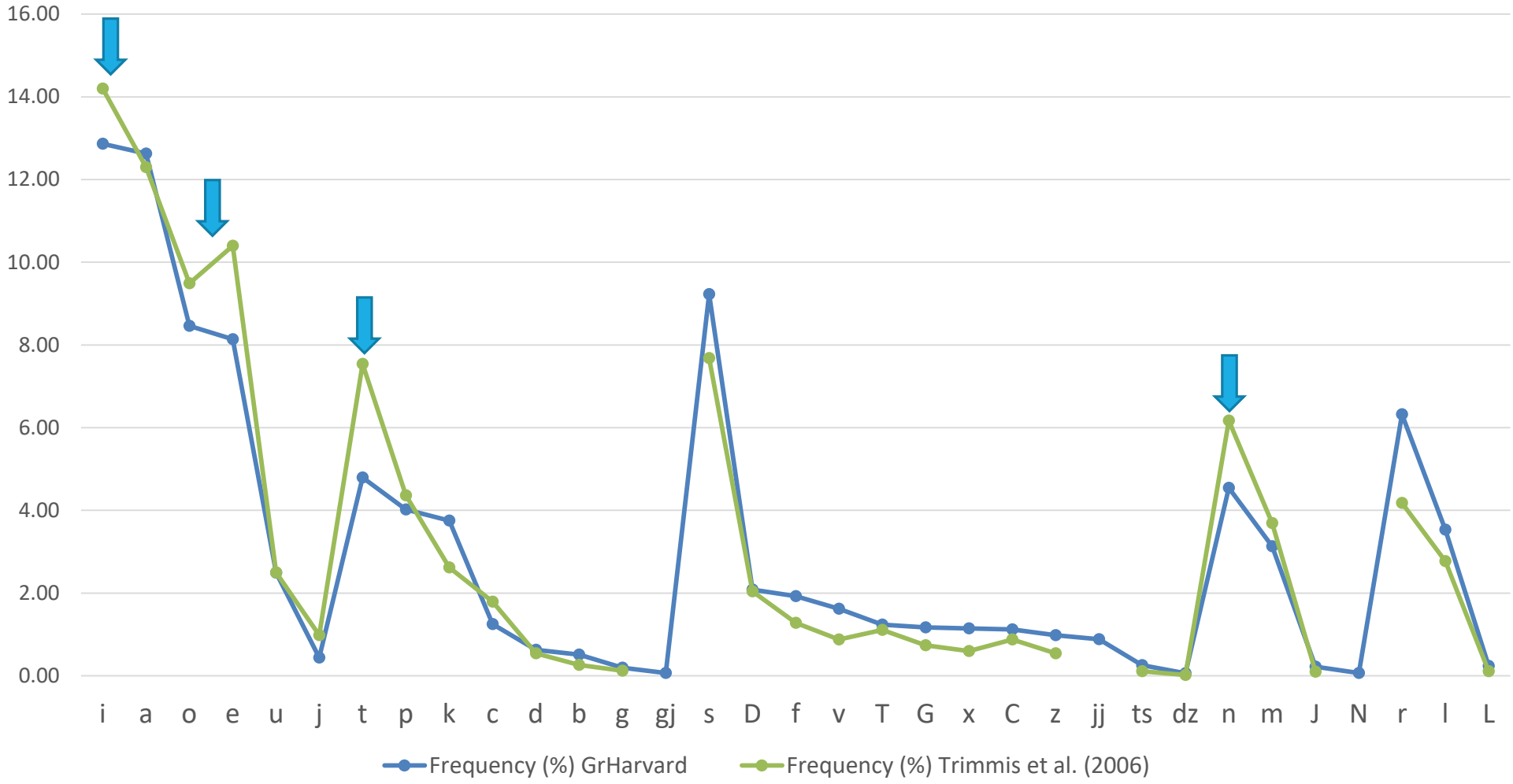
PHONEME FREQUENCY DISTRIBUTION

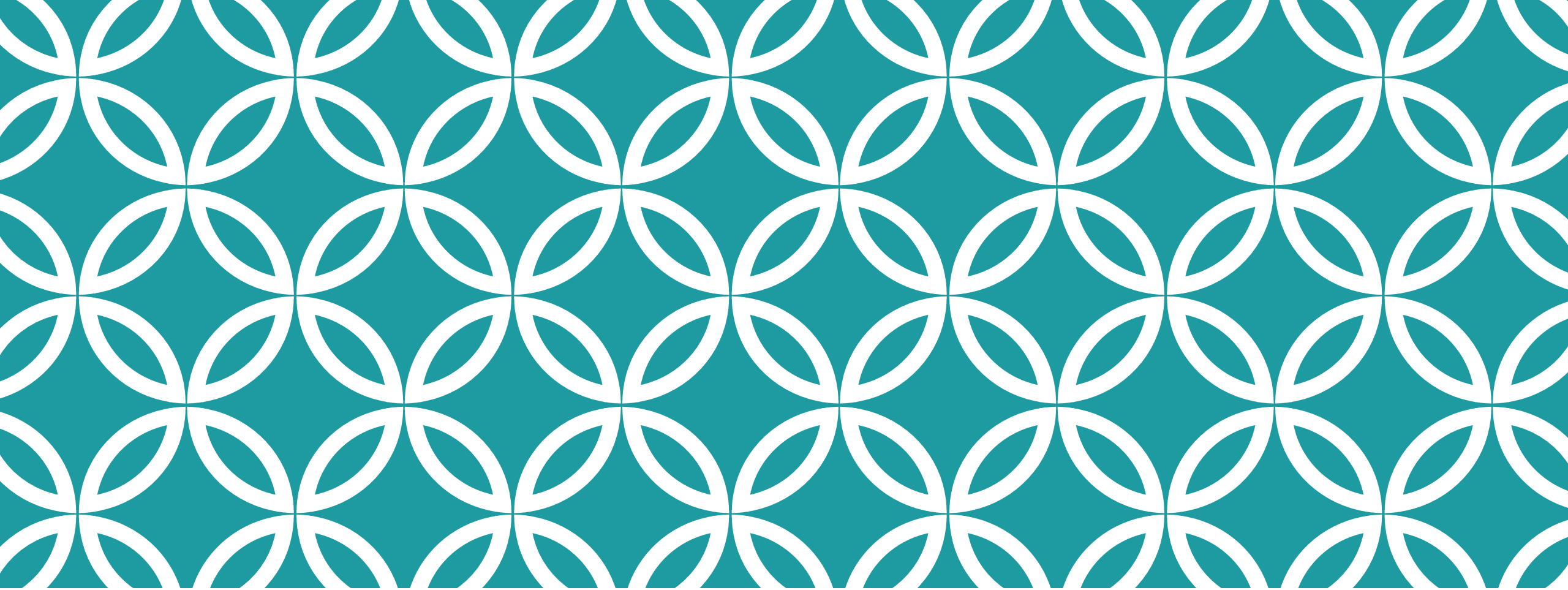
- Total number of phonemes in keywords: 20,206
- Comparison with IPLR
  - "C Corpus": 34 million tokens



PHONEME FREQUENCY DISTRIBUTION

- Comparison with Trimmis et al. (2006)
- Number of words analysed:
  - Trimmis et al.: 10,000
  - GrHarvard: 3,600 keywords (5,536 in total)

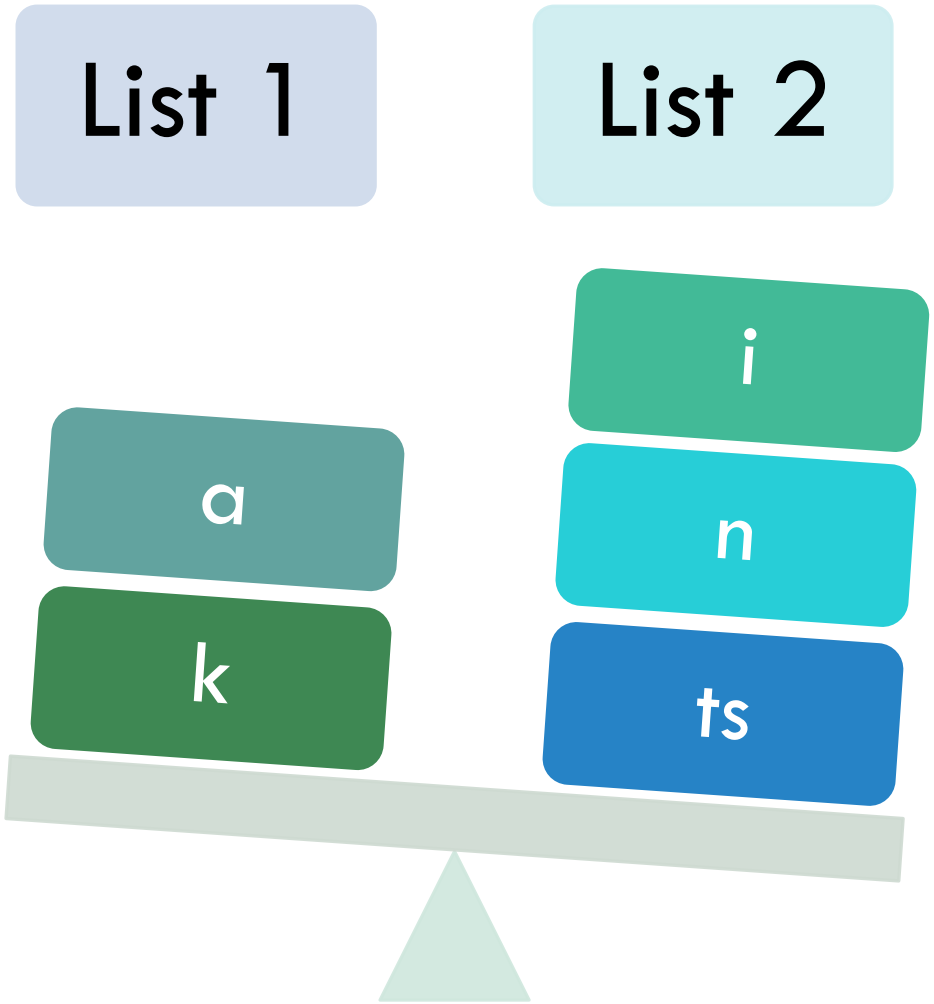




NEXT STEPS…

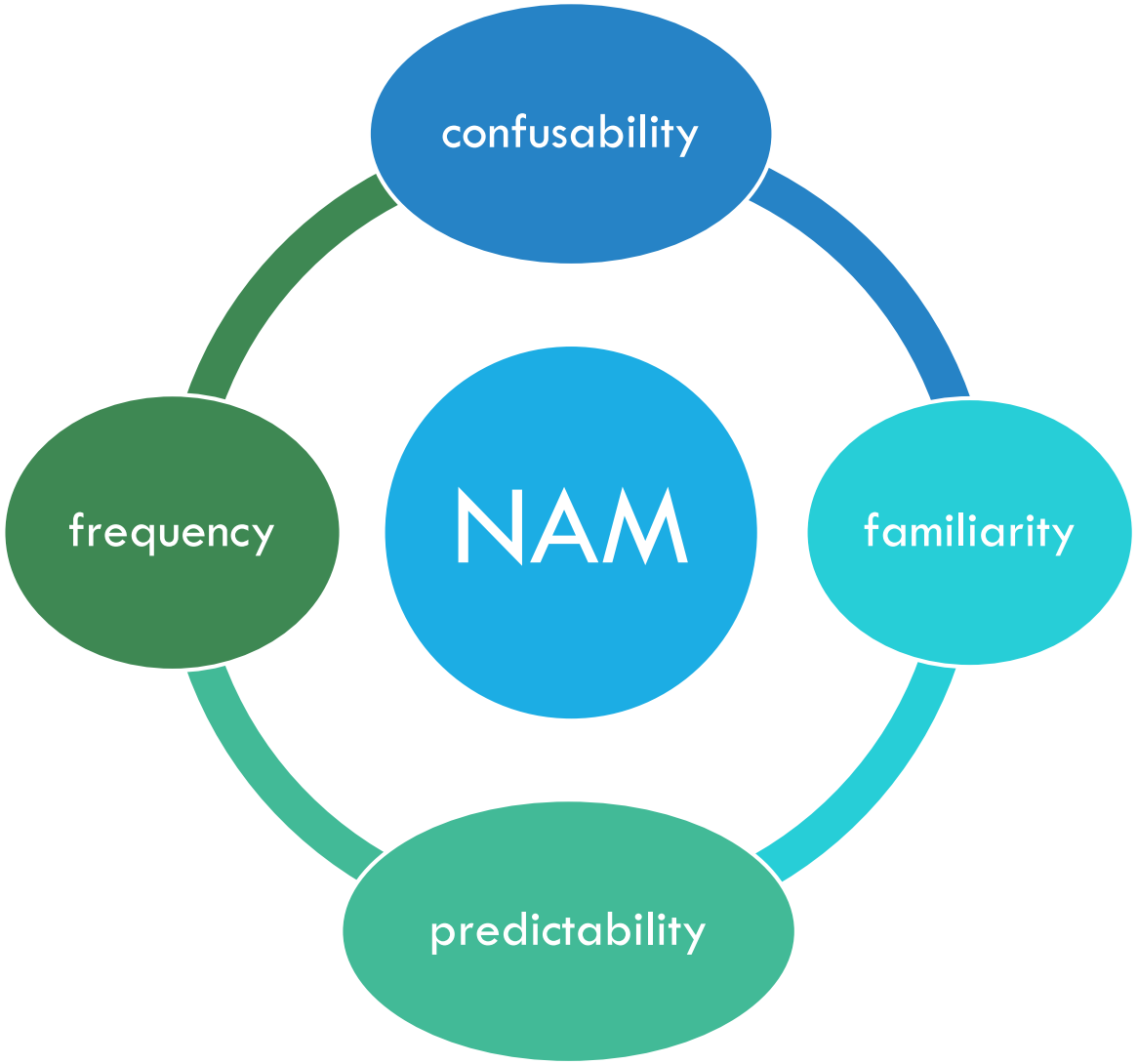
BALANCING THE CORPUS

- **Phonemic balance**
  - Some classes of sounds are more susceptible to masking by noise.
  - Phonetic content of corpus must be properly balanced to reflect distribution of speech sound classes that occur in language.
  - Lists must be phonemically equivalent.



BALANCING THE CORPUS

- **Neighbourhood Activation Model (NAM) (Luce, 1986)**
- Phonetically similar words in memory are organized for perceptual processing.
- Activation of set most consistent with acoustic-phonetic information in the speech waveform.
- Responses biased towards more frequent members.



BALANCING THE CORPUS

- **Keyword lexical confusability**
  - Words are not equally confusable → lexical or phonological neighbourhood density

  - Lexical candidates (Salasoo & Pisoni, 1985)

1. Acoustic-phonetic patterns
2. Semantic & syntactic information

BALANCING THE CORPUS

- **Keyword frequency**
- Hellenic National Corpus http://corpus.ilsp.gr
- Examples

Φόρεσε	το	**παλτό**	**προτού**	**βγεις**	**έξω.**	
0.0039		0.0021	0.0275	0.0022	0.2038	→ **0.0479** ‰
Τα	**χοντρά**	**βόδια**	**βόσκουν**	**φρέσκο**	**χόρτο.**	
	0.0038	0.0010	0.0006	0.0044	0.0017	→ **0.0023** ‰

BALANCING THE CORPUS

- **Keyword familiarity**
  - Familiarity ⇔ relative frequency in specified kinds of written or spoken language (Howes, 1957; Owens, 1961)
  - Familiarity test
    - 50 male & 50 female judges of different age groups
  - Effects of word frequency and familiarity are diminished when speech stimuli are inherently ill defined. (Salasoo & Pisoni, 1985)

BALANCING THE CORPUS

➢ Keyword predictability

➢ Higher intelligibility for words in **sentence context** vs. words in isolation (Miller et al., 1951)

➢ Intelligibility of words increases when the number of response alternatives decreases.

➢ The predictability of a word has an influence on its intelligibility. (Duffy & Giolas, 1974)

BALANCING THE CORPUS

- **Keyword predictability**
- Example

Μπάλωσε την **τρύπια τσέπη** με **βελόνα** και **κλωστή.**
[He/she mended the hole in the pocket with needle and thread.]

_____ την **τρύπια τσέπη** με **βελόνα** και **κλωστή.**

Μπάλωσε την _____ **τσέπη** με **βελόνα** και **κλωστή.**

Μπάλωσε την **τρύπια** _____ με **βελόνα** και **κλωστή.**

Μπάλωσε την **τρύπια τσέπη** με _____ και **κλωστή.**

Μπάλωσε την **τρύπια τσέπη** με **βελόνα** και _____.

BALANCING THE CORPUS

- **Keyword predictability**
- Examples

Μπάλωσε την **τρύπια** **τσέπη** με **βελόνα** και **κλωστή**.
HP **HP** **HP** **HP** **HP**

[He/she mended the hole in the pocket with needle and thread.]

Έκρυψε **μέσα** στον **τενεκέ** ένα **άχρωμο** **ζαφείρι.**

[He hid a colourless sapphire in the tin.]

BALANCING THE CORPUS

- **Keyword predictability**
- Examples

Μπάλωσε την **τρύπια** **τσέπη** με **βελόνα** και **κλωστή**.
HP **HP** **HP** **HP** **HP**

[He/she mended the hole in the pocket with needle and thread.]

Έκρυψε **μέσα** στον **τενεκέ** ένα **άχρωμο** **ζαφείρι**.
LP **HP** **LP** **LP** **LP**

[He hid a colourless sapphire in the tin.]

BALANCING THE CORPUS

- phoneme distribution
- frequency
- familiarity
- lexical confusability
- predictability

- total number of words per sentence
- number of syllables in keywords
- total number of syllables per sentence
- number of phonemes in keywords

lists with equally intelligible sentences

Prosodic parameters used as cues by the listener for understanding sentences (stress, word grouping)

BALANCING THE CORPUS

- **Optimizing selected parameters**
  - Automated optimization procedure
  - Interchanging sentences until lists are as equal as possible regarding selected parameters
- **Equating sentence difficulty**
- Listening tests (Nilsson et al., 1994)
  - present sentences in noise at a fixed SNR to normal-hearing listeners
  - measure percent intelligibility of keywords
  - change MS level of sentences according to intelligibility score
- Final word in sentence is less intelligible (Bell & Wilson, 2001)

RECORDINGS & LISTENING EXPERIMENTS

- Recordings with adults and children
- Noise masking with different SNR values
- Listening tests with native listeners
- Optimizing and finalizing corpus lists

- The GrHarvard Corpus is freely available to the research community.
- Contact: Anna Sfakianaki asfakianaki@csd.uoc.gr

EXAMPLE OF INTELLIGIBILITY TEST IN NOISE

- S1 -10dB -5dB 0dB white noise





- S2 -10dB -5dB 0dB babble noise





- S3 -10dB -5dB 0dB competing speaker





THANK YOU FOR YOUR ATTENTION!

(GA, Article 29): "This project has received funding from the EU's H2020 research and innovation programme under the MSCA GA 675324"

ENRICH EUROPEAN TRAINING NETWORK